

DOCUMENT RESUME

ED 472 767

SE 067 165

AUTHOR Webb, Norman L.
TITLE Evaluation of Systemic Reform in Mathematics and Science: Synthesis and Proceedings of the Annual NISE Forum (4th, Arlington, VA, February 1-2, 1999). Workshop Report.
INSTITUTION National Inst. for Science Education, Madison, WI.
SPONS AGENCY National Science Foundation, Arlington, VA.
REPORT NO No-8
PUB DATE 1999-12-00
NOTE 195p.
CONTRACT RED-9452971
AVAILABLE FROM National Institute for Science Education, University of Wisconsin-Madison, 1025 W. Johnson Street, Madison, WI 53706. Tel: 608-263-9250; Fax: 608-262-7428; e-mail: niseinfo@education.wisc.edu; Web site: <http://www.wcer.wisc.edu/nise/publications>.
PUB TYPE Collected Works - Proceedings (021) -- Reports - Descriptive (141)
EDRS PRICE EDRS Price MF01/PC08 Plus Postage.
DESCRIPTORS *Educational Change; Educational Policy; Elementary Secondary Education; Evaluation Research; Mathematics Education; Professional Development; Research Design; Research and Development; Science Education
IDENTIFIERS *National Institute for Science Education

ABSTRACT

This document reports on the National Institute for Science Education (NISE) forum on the Evaluation of Systemic Reform in Mathematics and Science. The purposes of the forum were to reflect upon what to understand about the evaluation of reform in the educational system, and to encourage and support continuing efforts to learn more about how evaluation can serve the multiple analytic needs in systemic reform for accountability, efficiency, and decision-making. The National Science Foundation's (NSF) six critical drivers describe the components of a successful systemic reform process: (1) an array of evidence that the reform has enhanced student performance in challenging mathematics and science material; (2) promotion of improved achievement by all students in the system; (3) implementation of a comprehensive, standard-based curriculum supported by needed professional development and assessment practices; (4) development of a coherent and consistent set of policies that support educational systemic reform; (5) convergence of all resources to support the systemic reform through a focused and unitary strategy; and (6) broad-based support from all segments of the community. (KHR)

Workshop Report No. 8

Evaluation of Systemic Reform in Mathematics and Science

Synthesis and Proceedings of the Fourth Annual NISE Forum

Norman L. Webb

 PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

P. White

 TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

 U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to
improve reproduction quality.

 • Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

 Funded by the
National Science Foundation

BEST COPY AVAILABLE

SEA 7465

National Institute for Science Education (NISE) Publications

The NISE issues papers to facilitate the exchange of ideas among the research and development community in science, mathematics, engineering, and technology (SMET) education and leading reformers of SMET education as found in schools, universities, and professional organizations across the country. NISE Occasional Papers provide comment and analysis on current issues in SMET education including SMET innovations and practices. The papers in the NISE Research Monograph series report findings of original research. The NISE Conference and Workshop Reports result from conferences, forums, and workshops sponsored by the NISE. In addition to these three publication series, the NISE publishes Briefs on a variety of SMET issues.

The Forum and its proceedings were supported by a cooperative agreement between the National Science Foundation and the University of Wisconsin-Madison (Cooperative Agreement No. RED-9452971). At UW-Madison, the National Institute for Science Education is housed in the Wisconsin Center for Education Research and is a collaborative effort of the School of Education, the College of Engineering, and the College of Letters and Science. The collaborative effort is also joined by the National Center for Improving Science Education, Washington, DC. Any opinions, findings, or conclusions are those of the author and do not necessarily reflect the view of the supporting agencies.

Workshop Report No. 8

**Evaluation of Systemic Reform in Mathematics and Science
Synthesis and Proceedings of the Fourth Annual NISE Forum**

Norman L. Webb

National Institute for Science Education
University of Wisconsin-Madison

December 1999

Acknowledgments

The author extends his thanks to Andrew Porter and Paula White for their thoughtful review of the first draft of this document. Thanks are also due Margaret Powell, editor, and Lynn Lund, secretary, who completed preparation of the document for publication.

Contents

Preface	v
Agenda	vii
Opening Keynote: Challenges to Evaluating Systemic Reform, <i>Norman L. Webb</i>	1
Panel I: Understanding Evaluation of Systemic Reform.....	9
The Detroit Urban Systemic Initiative: A Promising View of Systemic Reform, <i>Juanita Clay Chambers</i>	13
Understanding Evaluation of Systemic Reform: Purposes and Vision for Evaluation, <i>Daniel J. Heck</i>	27
Tracking the Theory of Change: A Moving Target, <i>Zoe A. Barley</i>	35
Evaluating Systemic Reform: A Complex Endeavor, <i>Iris R. Weiss</i>	47
Introduction to Breakout Session I Question Summary.....	53
Breakout Session I: Defining the Problems of Evaluating Systemic Reform	55
Panel II: Models and Approaches to Evaluation of Systemic Reform.....	57
Critical Elements of an Evaluation of Systemic Reform, <i>Patrick M. Shields</i> <i>Andrew A. Zucker, and Nancy E. Adelman</i>	61
Evaluators Roles: Walking the Line Between Judge and Consultant, <i>Jeanne Rose Century</i>	71
Assessing Student Outcomes, <i>Norma Dávila</i>	77
Understanding the Value of NSF's Investments in Systemic Reform, <i>Mark St. John</i>	92
Introduction to Breakout Session II Question Summary	111
Breakout Session II: Successful Strategies for Evaluating Systemic Reform	113
Panel III: Findings on Systemic Reform from Evaluation and Research	115
Discovering from <i>Discovery</i> : The Evaluation of Ohio's Systemic Initiative, <i>Jane Butler Kahle</i>	119
Evaluative Findings on Systemic Reform: Lessons Learned from NSF, <i>Daryl E. Chubin</i>	131
Value-Added Indicators, <i>Robert H. May</i>	135
Quantitative and Qualitative Data in the Theory of Systemic Reform, <i>William H. Clune</i>	145
Introduction to Breakout Session III Question Summary	149

Breakout Session III: Findings on Systemic Reform from Evaluation and Research . . .	151
Synthesis: 1999 NISE Forum.....	153
<i>Cora B. Marrett</i>	
Synthesis: 1999 NISE Forum.....	157
<i>Ernest House</i>	
Closing Observations	161
<i>Marshall Smith</i>	
Appendix A: Participant List.....	171
Appendix B: Fourth Annual Forum Evaluation	187

Preface

Norman L. Webb
University of Wisconsin-Madison

Ten years ago the National Council of Teachers of Mathematics released the first set of K-12 national content standards. Over the past decade, standards have been developed for most other content areas. Now nearly all of the states have content standards and assessments for mathematics, science, and language arts. The advancement of systemic reform has coincided with this massive effort on the part of states and districts to describe and assess more clearly what students should be able to know and to do in a multiplicity of content areas. Coinciding and closely linked with standards-based reforms, systemic reform has evolved from the theory developed by Smith and O'Day in 1991 into practice as a change strategy for surmounting the difficult problem of enabling all students to meet challenging content standards.

A national forum on evaluating systemic reform is both timely and necessary at this crucial point in the advancement of system-wide improvement. After a decade of experience, research studies, evaluations, and reflection, we have a considerable amount of information on attempts towards systemic reform and its evaluation. A spectrum of models of systemic reform that varies widely in the degree of success emerges from this information. The National Science Foundation (NSF) has spent hundreds of millions of dollars on systemic initiatives; now under pressure, Government Performance and Results Act (GPRA) personnel are seeking hard evidence of what the true impact of its massive effort to improve science and mathematics student performance has been. The National Institute for Science Education (NISE) Forum on the Evaluation of Systemic Reform in Mathematics and Science has two purposes. The first is for us to reflect on what we understand about the evaluation of reform in education systems. The second is to encourage and support continuing efforts to

learn more about how evaluation can serve the multiple analytic needs in systemic reform for accountability, efficiency, and decision-making. See Appendix B for a summary evaluation of the Forum-based evaluations completed by Forum participants

Our attention at this Forum and the work of NISE in studying systemic reform focuses on reform in mathematics and science. We acknowledge the important interactions of mathematics and science with other content areas and do not want the limiting of our focus to these two content areas to be interpreted as ignoring the value of other content areas. We have restricted our attention to mathematics and science because of the mission of the National Science Foundation and the benefits for studying reform with a content-specific approach. By attending to mathematics and science, we can build on the significant research that has been conducted on teaching and learning in these content areas. We can more easily trace activity through systems and find the connections among policy, administration, curriculum, and learning by focusing on these content areas. Systemic reform only in mathematics and science, however, is insufficient for full systemic reform. Thus, what we learn from evaluating systemic reform in mathematics and science will be relevant to the evaluation of related reform in any content area and to systemic reform in general.

A cornerstone of systemic reform is the establishment of high standards and a commonly shared vision or image of an idealized education system (Smith & O Day, 1991). More traditional reforms focus on a single component or unit and incremental change, whereas systemic reform considers all of the components, their interactions with each other, and their alignment in attaining common goals. In theory, school-based reform, curriculum reform, and other

singularly focused reform initiatives will be insufficient to sustain an effort to attain significant improvement in student learning without attending to other system components. Those successes that can be achieved through school-based reform will be deterred or inhibited by shifts in policy through state and district mandates or a diminishing teaching force of knowledgeable and well-trained teachers. Standards-based reform is important to a systemic reform, but does not imply that the reform is directed toward systemic change. Other components within the system, such as professional development, accountability, teacher preparation, and resource allocation, need to be addressed to achieve standards-based systemic reform. A state or district education system will make progress towards systemic reform when policies, administration, teaching, and curriculum are working in concert with each other in an effort directed toward promoting improved learning of challenging content by all students. The NSF's six critical drivers describe the components of a successful systemic reform process:

- An array of evidence that the reform has enhanced student performance in challenging mathematics and science material.
- Promotion of improved achievement by all students in the system.
- Implementation of a comprehensive, standards-based curriculum supported by needed professional development and assessment practices.

- Development of a coherent and consistent set of policies that supports educational systemic reform.
- Convergence of all resources to support the systemic reform through a focused and unitary strategy.
- Broad-based support from all segments of the community.

Over the past four years, the NISE systemic reform team has studied system reform and its evaluation. We have interacted on a number of occasions with those who were doing the evaluations of systemic initiatives and systemic reform. We have tried first to illuminate what the questions are that we should be asking about the evaluation of systemic reform. During our exploration of these issues, we mined the evaluation literature and talked to those who were trying to evaluate systemic reform. Then, we studied specific strategies and approaches for conducting evaluations of systemic reform in mathematics and science. Out of this work we have developed a basic understanding of the evaluation of systemic reform. That process continued at the 1999 Forum.

References

- Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman, & B. Malen (Eds.), *The politics of curriculum and testing (Politics of Education Association Yearbook, 1990)* (pp. 233-267). London: Taylor & Francis.

AGENDA

Fourth Annual NISE Forum: January 29, 1999

Evaluation of Systemic Reform in Mathematics and Science

Sunday, January 31

4:00-9:00 Registration

Monday, February 1

7:00-8:30 Registration

7:00-8:15 Continental Breakfast

8:30-9:30 **Welcomes, Overview**

Introduction to NISE and Charge for the 1999 FORUM, Andrew Porter,
NISE

Challenges to Evaluating Systemic Reform, Norman L. Webb, NISE

Need for Evaluation of Systemic Reform, Luther Williams, NSF*

Welcome to the FORUM, John B. Hunt, NSF

9:30-10:45 **Panel I: *Understanding Evaluation of Systemic Reform***

Chair and Discussant: Bernice Anderson, NSF

The Detroit Urban Systemic Initiative: A Promising View of Systemic
Reform, Juanita Clay Chambers, Detroit Public Schools

Understanding Evaluation of Systemic Reform: Purposes and Vision for
Evaluation, Daniel J. Heck, University of Illinois

Tracking the Theory of Change, A Moving Target, Zoe A. Barley,
Western Michigan University

Evaluating Systemic Reform: A Complex Endeavor, Iris R. Weiss,
Horizon Research

10:45-11:00 Refreshment Break

11:00-12:15 **Breakout Session I:**

Defining the Problems of Evaluating Systemic Reform

12:15-1:45 Lunch

1:45-3:00 **Panel II: *Models and Approaches to Evaluation of Systemic Reform***

Chair and Discussant: Larry Suter, NSF

Critical Elements of an Evaluation of Systemic Reform, Patrick M.
Shields, Andrew A. Zucker, and Nancy E. Adelman, Stanford

Research International (SRI)
Evaluators Roles: Walking the Line Between Judge and Consultant,
Jeanne Rose Century, Education Development Center
Assessing Student Outcomes, Norma Dávila, University of Puerto Rico
Understanding the Value of NSF's Investments in Systemic Reform,
Mark St. John, Inverness Research

- 3:00-3:15 Refreshment Break
- 3:15-4:30 **Breakout Session II:**
Successful Strategies for Evaluating Systemic Reform
- 4:30-6:00 **Reception**
Remarks by a representative for Vernon J. Ehlers, U.S. House of
Representatives (Michigan)

Tuesday, February 2,

- 7:00-8:15 Continental Breakfast
- 8:30-9:45 **Panel III: Findings About Systemic Reform from Evaluations and Research**
Chair and Discussant: Julio Lopez-Ferrao, NSF
Discovering from Discovery: The Evaluation of Ohio's Systemic
Initiative, Jane Butler Kahle, Miami University
Findings About Systemic Reform from Evaluations and Research,
Daryl E. Chubin, NSF
Value-Added Indicators, Robert H. Meyer, University of Wisconsin
Quantitative and Qualitative Data in the Theory of Systemic Reform,
William H. Clune, NISE
- 9:45-10:00 Refreshment Break
- 10:00-11:00 **Breakout Session III:**
Information Needs for Driving Future Systemic Reform
- 11:15-12:15 **Wrap Up**
Summary: Evaluation and Systemic Reform, Marshall Smith, U.S.
Department of Education
Conference Syntheses: Ernest House, University of Colorado;
Cora B. Marrett, University of Massachusetts-Amherst

*Unable to attend

CHALLENGES TO EVALUATING SYSTEMIC REFORM

Norman L. Webb
National Institute for Science Education

Welcome to this NISE Forum on Evaluating Systemic Reform. I am excited about the line-up of speakers and the diversity and the depth of experience all of you bring to this most important issue. Our conference format, which has evolved over a number of years, has worked well for enabling speakers to raise stimulating ideas that are discussed, dissected, and added to in the small-group discussions.

Evaluation of systemic reform is one of the crucial issues facing education today. It is like solving a giant jigsaw puzzle without the aid of the picture on the cover. Only through effective use of information and the close scrutiny of evaluation studies can improvements in our education systems be documented and understood. Without accurate and informative studies of the reform process, we face the prospect of repeating failures, acting without any sense of progress, and being subject to repeated whiplash from the onslaught of political and educational fads.

Systemic reform is one of the most innovative, massive, and ambitious attempts at education improvement our country has experienced since the curricula reform and Great Society era of the 1960s. The systemic initiatives of the National Science Foundation (NSF) have been a bold risk venture to improve science and mathematics education. They confront directly our society's needs for a strong economy and informed citizens. Congressman Vernon Ehlers, who will speak to us at the reception this afternoon, chaired a committee that in September released a Congressional report entitled, *Unlocking Our Future: Toward a New National Science Policy*. I recommend that all of you read this report, which is available on the Web. We also have copies on the display tables. Ehlers committee was charged with developing a long-range science and technology policy for the nation. This significant document

recognizes the vital role that education must play in this process. I quote from the report:

Our system of education, from kindergarten to research universities, must be strengthened. Our effectiveness in realizing the vision [to maintain and improve our country's pre-eminent position in science and technology] will be largely determined by the intellectual capital of the Nation. Education is critical to developing this resource.

Mathematics, science, and technology continue to advance at a rapid pace. For education to produce the intellectual capital to maintain our nation's economic strength requires that our schools do things they have never done before. Our schools are challenged to teach a more diverse population than ever before. They are required to teach somewhat different mathematics and science that has never before been taught on such a large scale. And, schools are asked to do this in a rapidly advancing technological environment.

Complacency can breed mediocrity. Ignoring the reservoir of untapped talent of the under-served in education, and having students be any less than they can be weighs down our society and creates lethargy and failure. Maya Angelou, in a recent address to the Wisconsin teachers, encouraged them to teach each youth as if she or he is the next Einstein, Andrew Wiles, Madame Curie, or Bill Gates. Gloria Ladson-Billings reminds us that what students can learn is not predetermined.

Ten years ago, systemic reform was a topic found only in books. We now have nearly a decade of learning and experience about how education systems have tried to advance large-scale change. These experiences have breathed life into this

theoretical vision toward change. Simply stated, systemic reform is

a *process* that extends over a long period of time and that has to engage a number of people in *system* improvement through *changing multiple system components* and *their interconnections* concurrently.

Systemic reform in education does not imply uniform practice nor the prevention of innovation. It does not imply only one strategy for change. Nor does it imply that there has to be a strong centralized system rather than a more locally controlled system. It does imply that a system needs to add greater stability, improve alignment, remove barriers and countervailing forces, create stronger links among components, and work with all teachers so that all students will have the chance to obtain knowledge of important science and mathematics.

Nobody said reform is easy. I draw strength from what Neil Postman wrote in the 1970s in his book, *Teaching as a Conserving Activity*. What makes education resistant to change by those who are well meaning and have the knowledge of what should be also inoculates education from the destructive viruses of fools and ill-placed quick fixes, and charlatans.

An important role of evaluation is to generate models and conceptualizations of what is being evaluated. Many of the evaluators present here have advanced their models of systems and systemic reform, including the SRI (Stanford Research Institute) pyramid to name one example. A very simple model of an education system consists of four general components-policy, management, programs, and student outcomes. These components and their functions do not reside at any one level such as the state level, school level, or classroom level, but incorporate and exist at all of these levels. Clearly, other components could be added to this simple model, such as the community.

What has distinguished systemic reform from other types of reforms is that other reforms have focused primarily on change in one and only one of the components.

Curriculum initiatives address the total program. School-based decision making primarily attends to management. State legislation that imposes a graduation requirement exists within the policy arena.

Many of the non-systemic reform theories of change are generally linear and uni-directional. [Slide 4] Change in policy effects a change in management that effects a change in curriculum and instruction that then is to result in improved student achievement.

Systemic reform is based on an assumption that the system components are interconnected, non-linear, complex, and adaptive. [Slide 5] Each of the components have an influence on all of the other components. As such, education systems are better represented more as an ecology than an assembly line. The conceptualization of the system and the approach to change has strong implications for the evaluation and study of the system. For example, in systemic reform, student achievement is not only an outcome variable, but is also both an input variable and a process variable.

As an outcome variable, levels of student achievement are specified goals and indicators of student learning. As an input variable, information on student achievement is used to make decisions about program, management, and policy. As a process variable, measures of student achievement communicate expectations, influence the enacted curriculum, are essential for system alignment, and define the performance gap between groups. The multiple roles of each system component have important implications for evaluation. It is not sufficient for an evaluator to monitor only student learning. Evaluators of systemic reform also need to understand how those in the system make decisions based on student achievement.

As I think about the evaluation of systemic reform, I think about **Umberto Eco's** essay on the theoretical possibility of creating a map of the empire on a scale of 1 to 1. The map has to represent each feature of the empire exactly, but cannot be placed over the empire being mapped because then the climate would be affected, causing a change in the terrain and forcing another change in

the map. Eco ends his essay with two concluding corollaries:

Every 1: 1 map always reproduces
the territory unfaithfully.
At the moment the map is realized,
the empire becomes unreproducible.

Our hope in evaluating systemic reform is not to represent everything faithfully. To do so would imply the lack of a dynamic system, or a record of what does not exist anymore. Our charge is to seek data on key attributes that will maximize the information we can use to understand the extent and quality of reform.

Evaluation of systemic reform involves practical research that calls upon multiple tools, strategies, and knowledge bases. Polarized positions, such as quantitative not qualitative, policy research not practical research, reform not traditional, understanding not drill, have no place in the evaluation of systemic reform. All of these facets have a role and need to be considered. All of these techniques and other techniques and views have to be considered *in context* and as *context*.

One important function for evaluation is to describe what is happening, what has happened, and what will happen. From SRI's evaluation of the state systemic initiative program, we have important descriptive information such as the five main implementation strategies used by state systemic initiatives (SIs). We also know that over one third of the middle grade mathematics and science teachers in the SI states had participated over the first four years of the effort.

A second important function for evaluation is to judge and to verify systemic reform's value as a reform strategy. Its value needs to be established in the context of at least three currencies—in relation to the theory of systemic reform, in relation to the goals of each system's reform, and in relation to alternative strategies. Does the reform create better alignment in the system? Has the gap in performance among groups been reduced while raising overall student achievement? Has professional development

improved the capacity of teachers to provide quality instruction in mathematics and science? Has systemic reform led to more significant and sustained change than would have been achieved through allocating all of the funds for reform to the purchase of new curriculum materials?

Evaluations of systemic reform need to judge the value and worth of reform in a larger context. Clearly, we seek evidence of improved student learning. We also need to seek other significant outcomes and payoffs related to investment in a risk venture. Sometimes payoffs come in the form of new inventions, transportable innovations, and system learning. A small investment of .1% of an education budget may not produce the targeted goals, but may result in a product such as Tang or Velcro. The recently released report on the New National Science Policy points to federally supported research on the molecular mechanisms of DNA, the so-called

blueprint of life, that led to recombinant DNA technology (gene splicing), which in turn spawned an entire industry. An important function of the evaluation of systemic reform is to identify the innovations, the spin-offs, and the learning that take place. One unexpected derivative of the Puerto Rico systemic initiative has been the development of the entrepreneurial function of schools. The schools, funded by the state systemic initiative for only a limited amount of time, had to develop management and marketing skills to seek continuing funding from other sources within the community. This is an important finding, but it also needs to be analyzed to determine whether it is a viable strategy that will prove, for example, to be a sustained source of resources for needed, on-going professional development of teachers. What is the evidence that schools as entrepreneurial enterprises will meet the challenges of reform while not taking away teachers' needed time and energy in performing the important function of educating students? Where is the balance and how do teachers and principals reach a suitable compromise?

A third function of evaluation is to explain. One value of science is to explain physical phenomena, the structure and the

compositions of galaxies, the genetic make-up of living creatures, and the interaction of people. As in physics, astronomy, biology, and psychology, one value of systemic reform evaluation is to explain how reform leads to a large percentage of students achieving challenging and high quality mathematics and science. Through clear explanations of the link between reform efforts, classroom practices, and student activities, we come to understand how reform contributes to improved student learning. One form of explanation is to isolate the primary reasons for an impact, thus connecting an effect to a particular cause. Another form of explanation is by eliminating alternative hypotheses and increasing confidence in conclusions and the cause of an event. Explaining why change has occurred through systemic reform evaluation is more of the latter than the former.

A fourth function of systemic reform evaluation is to offer, or participate in, recommended changes toward improvement in the direction and the nature of reform. Michael Scriven, in his *Hard-Won Lessons in Program Evaluation* (1993), does not fully accept this position and argues that evaluators should steer toward evaluative conclusions, but are generally in a weak position to offer recommendations. Understanding the logic of a large education system requires significant effort by an evaluator, or anyone else. When the longevity of district superintendents can be as little as two or three years, systemic initiative directors come and go with frequency, and NSF staff continually rotate, the evaluators of systemic initiatives have become the one constant. The evaluator who understands what is happening becomes a critical source of information and insight. I believe that not only should the evaluator at least participate in drawing up recommendations and setting new courses of action, but that the evaluator is ethically required to do so.

Evaluators of systemic reform in mathematics and science face numerous challenges. Many of the speakers at this Forum will address some of these both in their papers and in their presentations. It is the challenges that make evaluation of systemic

reform so interesting. The *size and complexity* of large district and state systems force us to break the problem into smaller parts and to conduct a series of coordinated studies. The *dynamism* of education systems forces evaluators to develop iterative plans that have to be periodically updated and refined, as more information is gathered and more knowledge about the system and the progress of reform are gained. The need for reform to saturate the system, to go to scale, and to leverage resources forces evaluators to consider the whole system as the unit of analysis. Evaluators need to understand what Jane Kahle has called the pressure points within the system. They also need to analyze and judge the viability of strategies and theories of change to address the full problem. Summary information, such as mean test scores, is insufficient as a basis for decisions on systemic reform. More detailed information is essential. An important challenge for evaluators is to gain access to what information is needed to understand reform. I am more convinced than ever, from our work with Milwaukee Public Schools, that embedded evaluation- where evaluators work interactively with those in the system in addressing problems-is necessary to gain a deep understanding of what the system is and what is necessary to seek reform. Finally, any study of systemic reform needs to consider the time frame within which significant change is to be attained. Forcing judgment on the significant progress of reform and change in student learning after only one or two years ignores entirely the complexity of education systems and what is required to penetrate them in order to achieve sustained improvement.

A small group of us have been engaged in writing a book on the evaluation of systemic reform. This group, which includes Dan Heck, Jeannie Rose Century, Norma Dávila, Eric Osthoff, and myself, has built on the work of several NISE Fellows and the many others who have shared their experiences with us. A centerpiece of the book is a section describing the nine attributes that are important in considering the evaluation of systemic reform. We have clustered these attributes as they

relate to systemic reform and in contrast to other types of reform. Five target attributes, those that are essential if systemic reform is to be achieved, are alignment, saturation, linkages, equity, and quality. Two enabling attributes represent features of the system that need to be changed to lay the foundation for systemic reform. Two other enabling attributes are capacity and sustainability. The explanatory attributes, incentives and trade-offs, help to explain or provide reasons for the advancement, or non-advancement, of systemic reform. Confronted with complex challenges, such collaboration is essential.

Andrew Wiles-as a case in point— accomplished one of the most astonishing intellectual endeavors in our time when, in 1994, he proved Fermat's Last Theorem. For seven years, he devoted himself to the solution of a problem that had evaded the grasp of the greatest mathematicians for 350 years. Wiles first read about Fermat's theorem when he was ten years old. At the age of 41, he accomplished his boyhood dream. In so doing, he demonstrated the importance of building on the work of others from all corners of the world. Raised in Cambridge, England, he did his work at Princeton University. Wiles used the work of two Japanese-Taniyama and Shimura-along with that of Galois (which Wiles studied when he was a teenager in France), the work of a University of California at Berkeley mathematician Ken Ribet, and many more. After three years of non-stop work, Wiles attended a major conference on elliptic equations, where he learned about a method first devised by Kolyvagin, a Russian

mathematician. This method proved to be an important key to developing the proof.

Those of you who have come to this NISE Forum on Evaluating Systemic Reform have an opportunity to interact and learn from those who have been at the center of systemic reform evaluation for several years. These are the people who have performed nearly all of the work in evaluating systemic reform, or who have worked with those who have led the effort to evaluate systemic reform.

An important goal of this Forum is to provide you and others an opportunity to learn from each other and to learn how others are thinking about the problem. Who knows, maybe you will hear about an approach that will be the key to how best to study systemic reform [Final Slide]-An approach that will help us better to describe, explain, judge, and recommend how the pieces of the puzzle fit together. Hopefully, though, none of you will write in the margin, I have a truly marvelous demonstration of evaluation of systemic reform, which this margin is too narrow to contain without providing a full disclosure of what you have discovered.

References

- Postman, N. (1979). *Teaching as a conserving activity*. New York: Delacorte Press.
- Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, A publication of the American Evaluation Association, William R. Shadish, Editor-in-Chief, Number 58, Summer 1993. San Francisco: Jossey-Bass Publishers.

OVERHEADS USED

Evaluation of Systemic Reform Confronting the Challenges NISE 1999 Forum

Norman Webb

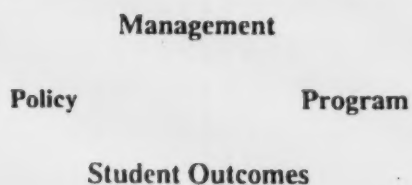
1

Systemic Reform is

- a process of
- system improvement through
- changing a multiple of components and
- their interconnections.

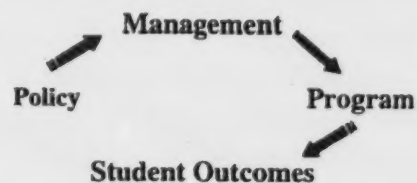
2

General Components of an Education System



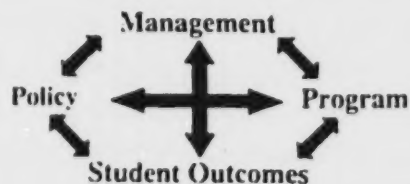
3

General Components of an Education System



4

General Components of an Education System



5

Evaluation of Systemic Reform

- Describe
- Judge
- Explain
- Recommend

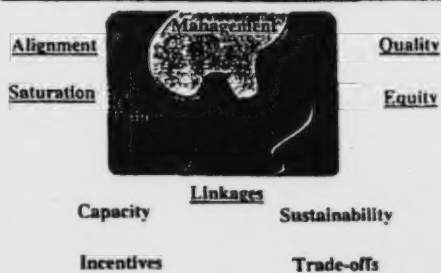
6

Evaluation Challenges

- Size and Complexity
- Dynamism
- Leveraging and scaling-up
- Access to information
- Time frame

7

Evaluation Framework



8

Evaluating Systemic Reform

1. Larger, more complex reform means larger, more complex evaluations.

9

2. You need to evaluate a reform strategy in light of other possible strategies.

10

3. It is technically difficult to measure impact on a "system."

11

4. It is politically difficult to wait to measure impact (e.g., on students) until you think it is reasonable to do so.

12

Panel I.: Understanding Evaluation of Systemic Reform

Panel Papers and Authors:

The Detroit Urban Systemic Initiative: A Promising View of Systemic Reform
Juanita Clay Chambers, Detroit Public Schools

Understanding Evaluation of Systemic Reform: Purposes and Vision for Evaluation
Daniel Heck, University of Illinois at Urbana-Champaign

Tracking the Theory of Change: A Moving Target
Zoe A. Barley, Western Michigan University

Evaluating systemic Reform: A Complex Endeavor
Iris R. Weiss, Horizon Research Inc.

Discussion Summary and Commentary: Understanding Evaluation of Systemic Reform

Norman L. Webb

Panel I set the stage for the other panels by providing a context for evaluating systemic reform. The four presenters discussed efforts towards systemic reform, reasons for engaging in the evaluation of systemic reform; conceptualizing a systemic initiative, and complexity involved in evaluating systemic reform.

Juanita Clay Chambers, a staff member of Detroit Public Schools and its Urban Systemic Initiative (DUSI), set the stage for discussing the evaluation of systemic reform. She described the challenges the Detroit Urban Systemic Initiative faces in the 10th largest school district in the country with limited resources, a high proportion of students from economically challenged families, a large and diverse group of teachers, and varying degrees of community support. Measuring the progress of the DUSI was viewed as important because the data challenged the initiative to consider the efficiency and effectiveness of its strategies. DUSI was designed to reach mathematics and science instruction in all schools and with all students in the district over five years. The school building was identified as the unit of change. A standards-based core curriculum for mathematics and science and professional

development in constructivist teaching were the main vehicles of reform. Two research questions drove the evaluation of the DUSI: Is the initiative effective in significantly improving student achievement in mathematics and science and is the initiative effective in producing a system that supports and sustains improved student achievement over time? A multi-method design was used for the evaluation. In-depth case studies were conducted in six schools, equally divided among the first two tiers of the three tiers of schools. Surveys were conducted in 54 schools randomly selected among the three tiers. All the mathematics and science teachers in each school sampled completed a survey. Findings from the evaluation supported an increase in the number of teachers engaged in professional development in mathematics and science; a higher number of students enrolled in advanced mathematics and science courses; and improved scores at all grade levels on the state criterion-referenced test and the district norm-referenced test.

Daniel Heck, at the time of his presentation a graduate student at the University of Illinois at Urbana-Champaign, outlined purposes for doing evaluations of systemic reform. He emphasized that systemic reform is a theory of change to move an education system toward an ambitious vision of learning. Evaluations of such ambitious efforts have to be equal to their task. Three developmental issues education systems face

define the meaningful purposes for evaluation of systemic reform: (1) the need to understand and manage change throughout a large, complex system; (2) the need to track the nature and extent of change over time; and (3) the need to build and test a theory of systemic reform operating in the system's context. Understanding and managing systemwide change requires evaluations to be approached from a systems view through adhering to the system in its totality, the complexity of the reform, and the full integrity of the reform. Because reforming education systems requires time, evaluations of systemic reforms need to consider a time frame over some duration. Instrumental in projecting progress over time are "baseline indicators" to establish the state of a system at some point in time, signs of progress, and trajectories towards future outcomes. All evaluations of systemic reform operate in a context that this approach to change is only a theory. Information produced through evaluations then becomes a major source for challenging or affirming this theory for change.

Zoe Barley, at the time a researcher and evaluator at Western Michigan University, emphasized one important reason for advancing systemic reform is the general failure of attempts to change specific system components in isolation such as teaching, instructional materials, and curriculum. She reported on the emerging understanding of theory-driven evaluation as an approach for shaping the evaluation of statewide systemic initiatives (SSIs) funded by the National Science Foundation. A first task in a theory-driven evaluation is to give form and specification to the theory. In the case of the SSIs this required documentation analyses and conversations with program directors. Two types of theories need to be explicated. A normative theory needs to be formed on the fit of the actual initiative activities with the intended intervention and the "design theory." A causative theory needs to be formed on what are the initiatives' impact.

In the initial stages of a systemic reform evaluation, one approach to presenting the design theory is to develop a logic model—a conceptual representation of the relationships

among the relevant inputs, intervening factors, intermediate benchmarks, and interventions with program staff. Through specific examples, Barley illustrated the usefulness of developing logic models to negotiate with project staff what an evaluation should emphasize and what are possible gaps between the planned work and the desired outcomes. She noted the need for rethinking the logic models based on untested theories. In one example, the role of the community in a large hierarchical urban district was, hypothesized as a necessary precursor to radical thinking of teaching and learning. This required the evaluators to rethink their logic model and the reallocation of evaluation resources. More effort was spent on observing the emerging relationship between key community leadership and the district administration. A theory-driven approach to systemic reform benefits both the evaluators in shaping their work and the reformers who gain a graphical representation of the relationship between their strategies, the multiple system factors, and the results they seek.

Iris Weiss, president of Horizon Research, Inc., draws a parallel between evaluation of systemic reform and other reform efforts from her experience in evaluating a number of SSIs. Although evaluators can use time-honored evaluation strategies with reasonable confidence in studying the reforms of individual system components, evaluators of systemic reform have less assurance. The increased complexity of evaluating multifacet systemic reform initiatives places evaluators in a position where proven techniques may not apply. They have to address and consider more components of the system; parts of the system not directly covered in the plan, sustainability of efforts, and expended resources. They also have to seek a broader and deeper understanding of educational systems. In more traditional efforts of reform such as professional development of teachers, the goals and the interventions generally are clearly identified. In systemic reform, although the goals are understood, the bewildering array of options for intervention increases the likelihood that

an evaluator's critique may only add to the confusion.

Evaluators of systemic reform are confronted with targeting the available resources and setting priorities for what will be studied. This requires negotiation with project leadership who will expect more than the evaluation can deliver. Evaluators need to seek a balance in reporting findings and making recommendations for a major redesign. Recommendations can be circumvented if reported confidentially to only a few people, can be meaningless because the bulk of the system prevents any mid-course corrections, or can exceed the

knowledge of what the existing leadership knows what to do. Evaluators have additional pressures exerted on them by funders who seek evidence of impact to report to policy makers long before the initiatives have had a reasonable time to surmount an effort necessary to make needed changes. Such pressures are increased by the lack of appropriate outcome measures of reform student achievement and system attributes, such as alignment. All of these reasons, along with high visibility within a charged political arena, distinguish evaluation of systemic reform from what most evaluators of more traditional and restricted programs face.

THE DETROIT URBAN SYSTEMIC INITIATIVE: A PROMISING VIEW OF SYSTEMIC REFORM

Juanita Clay Chambers
Detroit Public Schools

Statement of the Problem

The United States is currently experiencing fundamental changes in every aspect of society. The popular notion is to change the structure of an organization if it is not functioning properly. New structures are constantly being created without adequate attention to institutionalization of behaviors, if the desired changes are to occur. During the past decade, teachers and groups from the private sectors have advocated for fundamental changes in the structure and outcomes of public education (Goodlad, 1984). Frustration is high about the cost and quality of public education, with recent reports raising concern that our nation is still at risk, nearly a decade after the publication of "A Nation At Risk" pronounced the need for drastic changes in public education (Murnane & Raizen, 1988).

As widely reported in the media, there is a crisis in education with long-term social, economic, and political consequences for the future of the nation. Parents, educators, business leaders, university representatives, students, and the community in general cry out for school reform. As the 21st century approaches, the demand for change and improvement is heightened if students are expected to cope and live successfully in an ever changing technologically advanced society.

Across the nation, mathematics and science educators are engaged in large-scale reform efforts. A plethora of reports describe science achievement of the nation's youth (Jacobson & Doran, 1991; Mullins & Jenkins, 1988; National Commission on Excellence in Education, 1983) and have served as the driving force for change. Implementing reform in science education

requires teachers who are knowledgeable in science content, process and inquiry pedagogy (Radford, 1998). The challenges faced by urban districts are enormous and multifaceted. The National Science Foundation (NSF) has undertaken a national effort to respond to the problem in science education by undertaking comprehensive reforms through states and large urban districts with a high poverty index. NSF's strategy obligates states and large urban districts to mobilize broad-based coalitions to implement ambitious reform efforts in mathematics and science that are based on the premise that all children can learn if provided with a rich instructional environment. The second premise is that state and local policy changes can create these opportunities by providing a consistent and supportive policy structure for school improvement.

The Detroit Public Schools is poised on the brink of a new era that signals both the promise and the challenge of a fundamental transformation of mathematics and science education. The promise lies in the significant efforts that are presently underway upon which the Detroit Urban Systemic Initiative can build. The challenge is reflective of the significant needs that are inherent in any large urban district (district size and complexity; limited resources; numerous ongoing programs and initiatives to align with the overall reform effort; a high proportion of students from economically challenged families; a large and diverse group of teachers; historically disparate resources by building; varying degrees of community support involvement and empowerment). These challenges have often been allowed to overshadow the reservoir of intelligence, academic potential, curiosity and enthusiasm with which our students enter kindergarten.

The progress of the Detroit Urban Systemic Initiative (DUSI) is important to measure over time. By using measures in an ongoing assessment and improvement process, the initiative is challenged to consider its strategies in terms of the relative efficiency and effectiveness.

Process of Reform

DUSI is the district's main vehicle for achieving educational reform in mathematics and science. The initiative is linked to other essential components of the reform effort such as the Michigan Statewide Systemic Initiative (MSSI), the Detroit Mathematics and Science Centers, the Center for Learning Technologies, and the district Professional Development Council. As a result of the tight alignment of these components, a strong and holistic presence for mathematics and science education reform has been established in Detroit.

From the beginning, DUSI determined that changes made as a result of its work would be system-wide and of major consequence in totally reforming the teaching and learning of mathematics and science. Understanding the enormous challenges of a large urban district, DUSI developed a tiered process for implementation that allowed the district to learn from and scale up to full

implementation over the five-year USI grant period. The first tier of three constellations (33,195 students) began the process in 1994-95. Tier II followed the next school year, adding six constellations and three alternative schools (62,295 students). The third year, the final tier began the process with fourteen constellations and six alternative schools (72,5110 students). Thus, the scaling up process (outlined in Table 1) was planned to engage all schools beginning in 1996-97 and to have full implementation in all schools beginning in 1998-99.

Tier Structure for Scaling Up

DUSI is organized to enact its theory of reform through simultaneous change in all major systems of organization and structure, classroom practice (including curriculum, instruction, and assessment), professional development, and community involvement. This theory of reform was introduced to teachers through "Articulation Sessions" which were initiated as a constellation entered the first year of DUSI. These sessions brought together mathematics and science teachers from all schools in a constellation and served to open communication lines, foster cooperation between schools, provide staff development, and initiate partnering activities for students and teachers. Although

Table 1

	1993-94	1994-95	1995-96	1996-97	1997-98	1998-99
Briefing sessions, Targeted inservice	Tier I Tier II Tier III					
Develop awareness, Readiness and commitment	Tier I	Tier II				
Prepare for Action	Tier I	Tier II	Tier III			
Program Start up		Tier I	Tier II	Tier III		
Focus on complex thinking and interactive discourse			Tier I	Tier I Tier II		
Focus on problem based Instruction in real world contexts					Tier I Tier II Tier III	
Focus on reflection and performance assessment						Tier I Tier II Tier III

constellations served as the format for dissemination of the vision, the unit of change has been identified as the building. It is within the building that reform must be operationalized.

Using this communications vehicle, the new DPS standards-based curricula for mathematics and science were disseminated and further professional development and scale up for curriculum implementation was planned. At the elementary level, specialists in mathematics and science and building-level teacher leaders were trained in the new Core Curriculum to support other classroom teachers in mathematics and science instruction. At the secondary level, unit heads (middle school level) and department heads (high school level) were trained to assist mathematics and science teachers in constructivist teaching.

District support staff also were developed to serve as resources in curriculum content as well as pedagogy. Simultaneously, the district began to develop means to recognize, support, and enable the involvement of parents and other community members in the educational process.

Theoretical Framework

The Detroit Urban systemic Initiative (DUSI) has been structured to connect with classroom teachers in direct and strategic ways. One of the first activities at the onset of the DUSI involved creating a document which articulated principles of teaching and learning that might ultimately improve student understanding and achievement. This document, *A Constructivist Vision Towards Teaching, Learning, and Staff Development*, has served to inform administrators, teachers and staff of the DUSI vision for improvement by outlining the concepts and practices in a new approach to mathematics and science education. A key challenge in large urban districts is to help all stakeholders understand and work toward common goals. This constructivist vision document has served as a template for professional developers and school teams as they plan for future activities.

In order for DUSI reform efforts to succeed, it has not only been important for teachers to understand DUSI goals, but also for teachers to articulate their own ideas about teaching and learning and to think about changes that are needed for success. Several researchers support the idea that teacher beliefs are precursors to change and that the teacher is the crucial change agent in paving the way to reform (Ajzen & Fishbein, 1980; Crawley & Koballa, 1992; Cuban, 1979, Fullan & Miles, 1992; Jenlink, 1995).

Additionally, some researchers have noted that previous attempts at science reform fell short of successful change because they were not systemic in nature and often embodied a top-down model of change (Anderson & Mitchener, 1994; Bybee & DeBoer, 1994; Cuban, 1990; Fullan & Miles, 1992; Gordon, 1993; Sashkin & Egermeier, 1992).

A study by Haney, Czeniak, and Lumpe (1996) further articulated the importance of teacher beliefs on changes in practice:

In other words, teacher perceived outcomes regarding the behavior at hand and the likelihood that these outcomes will occur to be major influences on behavioral intention; therefore, contemporary reform cannot afford to ignore the importance of such beliefs. . . . The obstacles and enablers that the teachers were provided mattered less to them than did their beliefs about the positive and negative outcomes associated with the behavior. This finding suggests that teacher training should pay particular attention to the factors (such as providing curriculum materials, reducing class size, including flexible class scheduling, etc.) that are expected to lead to lasting changes in classroom practice. (p. 985)

Although targeting teacher belief systems may be viewed as critical to change, there are many other obstacles that may impede progress. Sparks (1994) made recommendations for effective, sustained, high quality staff development. Among the recommendations that were interwoven into

the design and format of the professional development experiences were:

- Keep the focus on student learning.
- Recognize that change affects staff members in personal ways.
- Change the organization's culture at the same time that individual teachers and administrators are acquiring new knowledge and skills.
- Use a systems approach to change.
- Apply what is known about the change processes to the improvement effort.
- Make certain that the learning process for teachers model the type of instruction that is desired.
- Provide generous amounts of time for collaborative work and various learning activities.

Evaluation

An evaluation process continues to examine the impact of the Detroit Urban Systemic Initiative on the improvement of mathematics and science in Detroit Public Schools. Emphasis is placed upon the implementation of standards that articulate what is important for students to know and do in mathematics and science; improved delivery systems, professional development, student enhancement activities, parental involvement and policy alignment. The major research questions for the evaluation follow:

1. Is the initiative effective in significantly improving student achievement and accomplishments in mathematics and science?
2. Is the initiative effective in producing a system supporting such improved student achievement and capable of sustaining this accomplishment over time.

Data were gathered from two major stakeholder groups (teachers and students) over the years comparing the students and teachers before and after they experienced the program.

Methodology

The Target Population

For two years, six schools were randomly selected for an in-depth study. This selection was based on the schools' position in the staging process of the initiative. K-12 systemic reform in Detroit will see students, K-12, and all teachers with responsibility for mathematics and science impacted by the change strategies as described. The goals of the initiative are: (1) to improve the mathematical and scientific literacy of all students; (2) to provide the mathematics and science fundamentals that will enable successful participation in a technological society; and, (3) to significantly increase the number of students that will enter mathematics, science and engineering careers.

Detroit paced the implementation of the objectives of the USI over the first three years by involving sets of the District's 23 constellations each year in stages. These K-12 constellations will become learning communities in which staff come together periodically to plan, train, and make articulation decisions toward achieving the goals of the initiative. While all constellations were involved in the change process from the beginning, they were at different stages depending on their current status as it relates to staffing patterns. A profile of each K-12 constellation was developed to determine its status in relation to the following three stages of involvement: (1) In-depth, successful implementation of strategies and identified professional development; (2) preparedness for implementation; and, (3) awareness/readiness/commitment. During the first year, three constellations were selected to start stage one, as their profiles revealed a significant number of activities in place for the purposes of the USI (these constellations became known as Tier I). In addition, during the first year, six other constellations began stage two, preparing for implementation (Tier II); and the remaining fourteen constellations began stage three developing awareness, readiness, and commitment (Tier III). For the second year of the initiative, the six

constellations that started in Tier II in the first year, began stage one, in-depth implementation; and the remaining fourteen constellations began stage two preparing for implementation. All constellations were involved with the in-depth implementation by the third year of the initiative. By the fifth year, the results of systemic change will be evident.

Of the six schools selected for this in-depth study, three schools are from Tier I, which was impacted more by the innovation and professional development offered by the initiative. The remaining three schools were selected from Tier II, which allowed for an investigation of the impact of the curriculum innovations, professional development, and curriculum implementation of the initiative. The second group of case study schools were selected because of exemplary performance on standardized measures.

In addition to the in-depth case studies, surveys were conducted in 54 schools randomly selected by tier. The sampling plan was a non-proportional sample to represent tiers and school levels. The sampling plan provides for random selection and reasonably sized samples. A two-stage sampling process was used for teachers with random selection of schools by tier and subsequent surveying of all mathematics and science teachers within the randomly selected school. Adequate representation by Tier (I, II, and III) was provided for. The sampling is disproportionate and appropriate weighting was conducted. All mathematics and science teachers in the randomly selected schools were surveyed along with their students. Focus groups were conducted among Tier I and II teachers, parents, Unit Heads/Department Heads.

The study follows an Institutional Cycle Design (Payne, 1994), where a group is first assigned to a treatment (Tier I) and then is tested. The second group (Tier II) would be tested at the same time as the first group and then exposed to the treatment. They would then be post-tested. Then a third group (Tier III) would be tested at the same time as the group two post-test and receive the treatment. They would be post-tested after receiving the

treatment. Program impact will be measured by Tier I post- versus Tier II pre-; Tier II post- versus Tier III pre-; and Tier III pre- versus Tier III post-tests. Data from the surveys were analyzed over a three-year period.

Summary of Progress

Some of the most important findings are:

- Increases were observed in the number of teachers engaged in mathematics and science professional development (PD) through the Mathematics and Science Centers and DUSI. In addition to formal PD, a comprehensive structure supporting change in teaching practices, which is not captured in the data, occurs on a daily basis in schools and classrooms. These offerings are tightly aligned with the desired changes in curriculum and pedagogy so that support for systemic reform is built into the process for continually upgrading teachers' skills.
- The Tier system for implementation has demonstrated DUSI's impact on both student and teacher measures. Tier I teachers reported increased use of standards-based instructional practices, more involvement of parents and community members in the mathematics and science programs, and greater teacher confidence in their ability to implement the standards-based instruction as a result of their involvement in PD. Students in Tier I schools reported more positive attitudes toward mathematics and science instruction and confirmed the increased use by teachers of standards-based instructional strategies.
- Teachers report good alignment of curriculum with their instructional practices, including emphasis on developing students' problem-solving

skills, abilities to make connections to the real world, skills and knowledge for excelling on local and national tests of mathematics and science achievement.

- The DUSI Summer Institute was a very successful staff development activity that resulted in improved teaching strategies, increased use of new mathematics and science curricula in the District, and building-level action plans. As a result of the PD program, teachers reported a high degree of confidence in their ability to implement the new standards-based curriculum and related teaching practices.
- An increasing percentage of students are becoming engaged in mathematics and science. Of note are the large increases in student enrollment in advanced courses. Also, while student programming has not changed greatly, many more students are involved in the program across all levels.
- Student performance results indicate a steadily-improving rate of achievement as measured by the state-required criterion-referenced test (the MEAP) and the district-required norm-referenced test (the Metropolitan Achievement Test) in both mathematics and science at all grade levels.

References

- Anderson, B. L. (1993). The stages of systemic change. *Educational Leadership*, 51(1), 14-7.
- Anderson, R. D., & Mitchener, C. P. (1994). Research on science teacher education. In D.L. Gabel (Ed.), *Handbook of research on science teaching and learning*. New York: Macmillan.
- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting*

social behavior. Englewood Cliffs, NJ: Prentice Hall.

- Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis*, 12 (3), 331-8.
- Crawley, F. E., & Koballa, T. R. (1992). *Attitude/behavior change in science education: Part I. Models and methods*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Boston, MA.
- Cuban, L. (1990). Reforming again, again, and again. *Educational Researcher*, 19, 3-13.
- Corcoran, T. B. (1995). Helping teachers teach well: Transforming professional development for teachers: A guide for state policymakers. *CPRE Policy Briefs: Reporting On Issues and Research in Education Policy*. New Brunswick, NJ: Rutgers, The State University of New Jersey, Carriage House at the Institute of Politics.
- Dewey, J. (1938). *Experience and education*. New York: Macmillan.
- Education Commission of the States. (1992). Introduction to systemic education reform. *The Restructuring of the Education System Series*, 2-6. Author.
- Finn, C. E. (1990). Professional development: The biggest reform of all. *Phi Delta Kappan*, 71 (8) 584-592.
- Fullan, M. (1996). Turning systemic thinking on its head. *Phi Delta Kappan*, 77 (6), 420-423.
- Fullan, M. (1993). Innovation, reform, and restructuring strategies. Association for Supervision and Curriculum Development. *Challenges and Achievements of American Education*, 116-133.
- Fullan, M. (1993). Change forces: Probing the depths of educational reform. London: The Palmer Press.
- Gordon, R. (1993). The irrational science of educational reform. Paper presented at the Annual Meeting of the *American Research Journal of Science Teacher Education*, 6 (4), 187-196.
- Kuhn, T. S. (1970). *The structure of scientific*

- revolutions* (2nd ed.). Chicago: The University of Chicago Press.
- Murnane, R. J., & Raizen, S. A. (Eds.). (1988). *Improving indicators of the quality of science and mathematics education in grades K-12*. Washington, DC: National Academy Press.
- National Commission on Excellence in Education. (1983). *A nation at risk*. Washington, DC: U.S. Government Printing Office.
- National Science Board Commission on Pre-College Education in Mathematics, Science, and Technology. (1983). *Educating Americans for the 21st century*. Report to the American People. Washington, DC: National Science Board.
- Sparks, D. (1994). A new form of staff development is essential to high school reform. *The Education Forum*, 60 (3), 260-266.
- Stein, M., Edwards, T., Norman, J., Roberts, S., Sales, J., Alec, R., & Clay-Chambers, J. (1994). *A constructivist vision for teaching and learning and staff development*. (ERIC Document #383-557).
- Tye, K. A. (1992). Restructuring our schools beyond the rhetoric. *Phi Delta Kappan*, 74 (1).

OVERHEADS USED

**Evaluation of Systemic Reform in
Mathematics and Science**



National Institute for Science Education
Detroit Urban Systemic Initiative
Eddie L. Green, Ed.D.
General Superintendents & Principal Investigator
Juanita Clay-Chambers
Associate Superintendents & Project Director

NISE/Phi 99 Detroit Public Schools

1

PROCESS OF REFORM

DUSI:
The Unitary vehicle for achieving
educational reform in mathematics
and science in the Detroit Public
School District.

NISE/Phi 99 Detroit Public Schools

2

**DUSI REFORM PROCESS
ADDRESSES MAJOR VARIABLES**

- District size and complexity
- Numerous ongoing programs and initiatives to align with the overall reform effort
- A high proportion of students from economically challenged families
- Historically disparate building resources
- A large and diverse group of teachers
- Varying degrees of community support, involvement, and empowerment

NISE/Phi 99 Detroit Public Schools

3

**VALUES WHICH DEFINE THE
CULTURE OF
SUCCESSFUL URBAN SCHOOLS**

- A pervasive belief that All students can learn and achieve at high levels
- Acceptance of the premise that schools must be a learner-centered, caring community
- A primary and central focus of professional staff on student Outcomes
- A consensus that "Everyone" is responsible for learning

NISE/Phi 99 Detroit Public Schools

4

FOUR CORNERSTONE INITIATIVES

- **Exit Skills:** Grade level Performance Standards to assure that individual students meet specific academic targets
- **Resource Coordinating teams:** Professionals (consisting of teachers, nurses, social workers, counselors, psychologists, attendance officer and other supporting agencies.) configured to address the barriers to learning

NISE/Phi 99 Detroit Public Schools

5a

FOUR CORNERSTONE INITIATIVES
(Continued)

- **K-12 Constellations:** 20 learner centered communities with neighborhood resources focused on student development and progress which create learning "villages" for learning and efficiency
- **Site Based Management:** A local governance involving school-community stakeholders which allows for decision making that is specific for a particular school. This involves the creation of a local council of administrators, teachers, parents and others concerned about student progress

NISE/Phi 99 Detroit Public Schools

5b

DUSI KEYS TO SUCCESS

- A commitment to student outcomes and metrics to gauge project progress
- Early strategic planning and capacity building
- Establishing and gaining commitment to a clear vision
- The development of a visionary guide entitled:
"A constructivist Vision for Teaching, Learning, and Staff Development"

HSE/Feb 99

District Public Schools

6

POLICY CHANGES

- Curriculum alignment with national and state standards
- Alignment of assessment with instructional practice
- Increased opportunities for content-specific professional development
- Initiated "new" delivery systems
- Encouraged greater parent and community involvement
- K-12 Mathematics and Science Resolution
- Increased graduation requirements in mathematics and science

HSE/Feb 99

District Public Schools

7

MANAGEMENT CHANGES

- Realigned Organizational Structure
- Project Director elevated to cabinet level status

HSE/Feb 99

District Public Schools

8

ARTICULATION OF THE VISION

- The vision becomes real in the classroom, in the school and constellation and across the system.
- At each of these levels, an expression of the vision is supported by programming efforts.

HSE/Feb 99

District Public Schools

9

CRITICAL ISSUES RELATED TO DUSI REFORM

- Creation of a "Learner-Centered" school Service Support Intervention Program which promotes Academic Excellence and High Achievement
- Systemic approaches to reduce/eliminate the "Barriers to Learning"

HSE/Feb 99

District Public Schools

10

DATA COLLECTION AND UTILIZATION

- Monthly Department/Unit Head updates
- Data Collection Task Force established
- Documentation notebooks
- Ninth Grade Restructuring Evaluation
- MEAP/MAT Data for decision making
- Professional Development Database
- Case studies & Services Rendered forms

HSE/Feb 99

District Public Schools

11

IMPACT OF DUSI

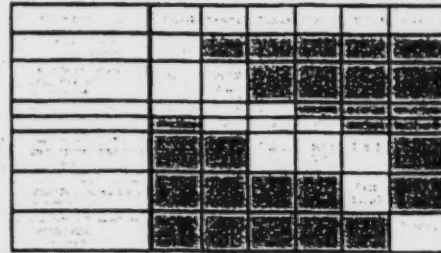
- Continuous improvement of test scores on standardized measures
- Increased student participation in Mathematics and Science Opportunities
- Improved delivery of standards-based instruction as a result of exciting professional development opportunities

HSE/Feb 99

District Public Schools

12

SCALING UP REFORM



HSE/Feb 99

District Public Schools

13

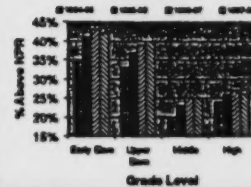
DUSI - PERFORMANCE OUTCOMES DATA

HSE/Feb 99

District Public Schools

14

MAT PERFORMANCE Mathematics

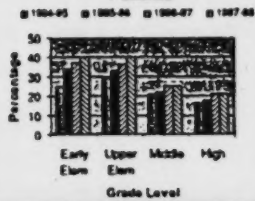


HSE/Feb 99

District Public Schools

15

MAT PERFORMANCE Science

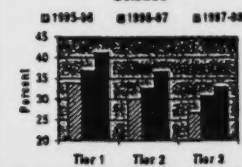


HSE/Feb 99

District Public Schools

16

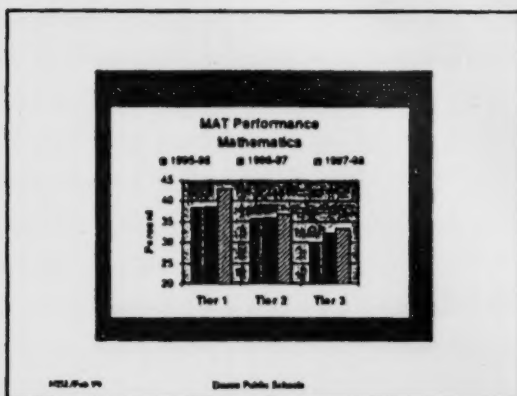
MAT Performance Science



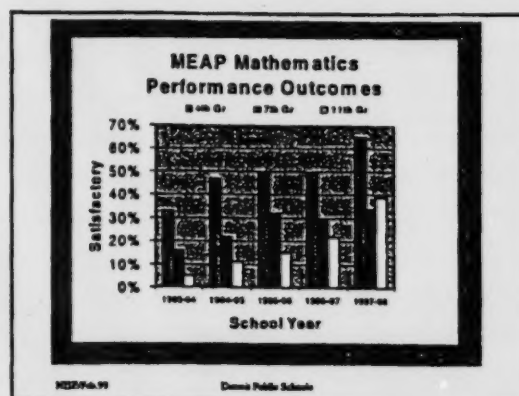
HSE/Feb 99

District Public Schools

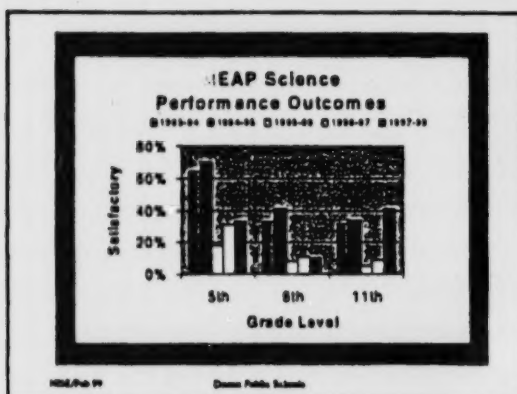
17



18



19



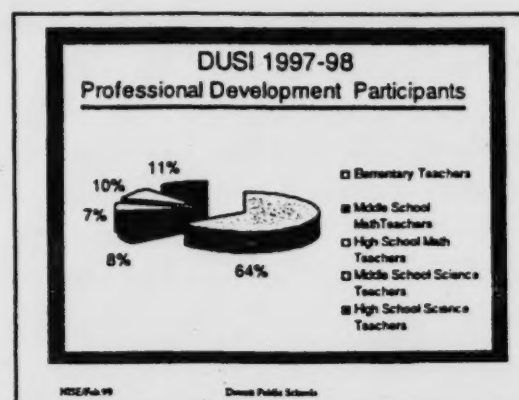
20

- TEACHER IMPACT**
- 100% of the teachers of the Detroit Public Schools impacted by vision
 - 75% of the teachers of the Detroit Public Schools receiving direct services
 - 85% of the teachers impacted by the DUSI Summer Institute for Professional Development
- HSE/Feb 99 Detroit Public Schools

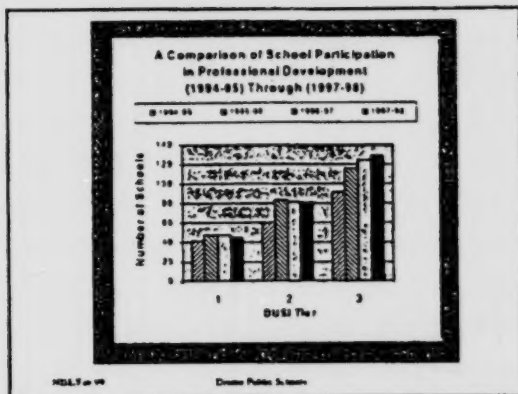
21

- INNOVATIONS IN PROFESSIONAL DEVELOPMENT**
- Release time for planning
 - Constructivist Workshops
 - Learning Logs
 - Connected Math Inservice
 - Algebra and Geometry Courses
 - Technology Strand Development
 - Ninth Grade Restructuring
 - Building PD Planning
 - Modeling
- HSE/Feb 99 Detroit Public Schools

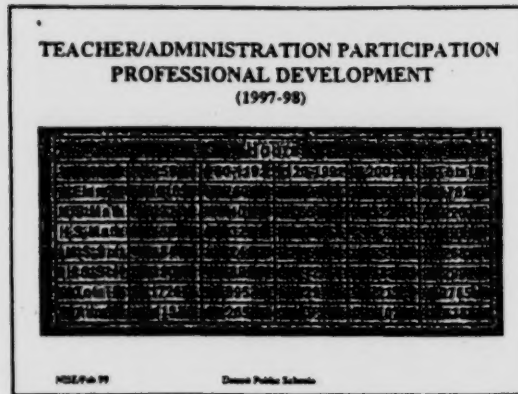
22



23



24



25

DUSI INNOVATIONS AND MAJOR DELIVERY SYSTEMS

- TERC Investigations
- Connected Mathematics
- Core Plus Mathematics
- 24 Challenge Mathematics
- FOSS
- Project AIMS
- Insights
- Project SEED
- VideoDiscovery Science Sleuths Videolaser Discs
- VideoDiscovery Bioforums Videolaser Discs

HSB/Feb 99 Detroit Public Schools

26

DUSI INNOVATIONS AND MAJOR DELIVERY SYSTEMS

- Model-It
- Technology (Pasco Probes)
- Project-Based Science
- Constructivist Teaching and Learning Practices
- DUSI Summer Institutes
- Family Math and Science
- Study Groups
- Center for Learning Technologies
- University Coursework
- Integrated Natural Science (District Developed Course)

HSB/Feb 99 Detroit Public Schools

27

DUSI INNOVATIONS AND MAJOR DELIVERY SYSTEMS

- Biology, Chemistry, Physics Support Series
- Apprenticeship Programs Henry Ford Hospital Karmans Cancer Center Wayne State University
- Science Fair Support
- Problem Based Instruction
- Center for Molecular and Cellular Toxicology (Wayne State University)
- Science in the City (Michigan State University)

HSB/Feb 99 Detroit Public Schools

28

Detroit Public Schools

For further Information Contact:
 Juanita Clay-Chambers,
 Associate Superintendent
 5057 Woodward Ave.
 Detroit, Michigan 48202

Phone: 313-494-1082
 Fax: 313-494-7864
 E-mail: juanita_chambers@dpsnet.detspub.k12.mi.us
 Web site: www.dpsnet.detspub.k12.mi.us

HSB/Feb 99 Detroit Public Schools

29

BEST COPY AVAILABLE

25

UNDERSTANDING EVALUATION OF SYSTEMIC REFORM: PURPOSES AND VISION

Daniel J. Heck

University of Illinois at Urbana-Champaign

Evaluation and Systemic Reform

Systemic reform is a theory of change intended to move an education system toward an ambitious vision of learning. Evaluation should inform and enhance its potential to do so (Rowland, 1994). It is troubling, then, that Carol Weiss (1991, pp. 223-224) described program evaluation as a political act that "tends to ignore the social and institutional structures within which the problems of the target group are generated and sustained." Weiss added that "most of the political implications of evaluation have an establishment orientation. They accept and bolster the status quo. They take for granted the world as defined in the existing social structure." Any evaluation that ignores the influence of existing educational structures and fails to look beyond traditionally accepted solutions will hardly be well-matched to systemic reform. Evaluations of systemic reform cannot behave as Weiss suggests many evaluations do. Systemic reform demands approaches to evaluation that match the considerable extent of the reform; the evolving goals of reform; the interdependent, emergent and responsive events and understandings of the reform; and the shifting political influences surrounding the reform (Bruckerhoff, 1997; Jenness & Barley, 1995; NSF, 1993; Ridgway, 1998).

Purposes for Evaluation of Systemic Reform

For what purposes do we evaluate systemic reform? Adaptation and innovation of activities and structures aligned toward a common, ambitious vision of learning characterize systemic reform in education.

Cronbach and colleagues (1980, p. 156-157) wrote that "evaluation at its best assists in a smooth accommodation of social activities and structures to changing conditions and ideals." In order for evaluations to "assist in a smooth accommodation" of systemic reform in an education system, evaluations must make sense of change on a large scale, over a long time, and within an evolving theory of reform. Michael Patton has offered one approach toward such an evaluation. Patton (1994) coined the term "developmental evaluation" to describe evaluation activity that focuses on continual innovation and adaptation such that versatile initiatives remain best suited to changing conditions and contexts. Development—that is, learning, innovation, and change—should be at the heart of systemic reform and its evaluation.

Education systems engaged in systemic reform of mathematics and science face three key developmental issues: (1) the need to understand and manage change throughout a large, complex system, (2) the need to track the nature and extent of change over time, and (3) the need to build and test a theory of systemic reform operating in the system's context. These three major issues render three meaningful purposes for evaluation of systemic reform. Each can be met well through a developmental evaluation focused on learning, innovation, and change. The first two of these purposes will be discussed here.

Understanding and Managing Systemwide Change

A key concept in systemic reform is systemwide change. The term systemwide is commonly used to describe the collection of districts and schools comprising an education

system. In systemic reform, systemwide also refers to the major functions of the education system-policy and governance, management and administration, instruction and learning. In the context of systemic reform, learning, innovation, and change develop interdependently across districts and schools and throughout the functions of the system. Evaluation of systemic reform should promote understanding and management of change developing across and throughout a system. A proposed developmental evaluation framework for understanding and managing systemwide change involves three closely related, but distinct, perspectives-a whole system view, a holistic view, and a systems view.

First, through a *whole system view* evaluation should provide sound and thorough descriptions of the education system, the reform, and their intersections. In order to facilitate understanding and management of systemwide change, evaluations should identify the districts, schools, and other structures along with the functions that comprise the education system. Evaluations should also identify the components of the reform effort, and most importantly, the specific parts of the system being targeted by components of the reform. The evaluation should highlight intended and unintended points of pressure and influence between the reform and the system. Evaluation audiences, who must understand or manage systemwide change, should maintain this whole system focus so that the big picture of reform is not lost in the details of planning or implementation (Bruckerhoff, 1997; Heck & Webb, 1996). The whole system view represents the *totality* of the systemic reform.

Second, evaluation should embed the multiple and dynamic objectives and activities of the reform within a *holistic view* of the reform (LeMahieu, 1997). In order to understand or manage systemwide change, evaluation audiences must appreciate how each activity or objective of the reform relates to the primary goals of the reform and overall vision of systemwide change. Moreover, when the evaluation focuses on specific districts or schools, or individual functions of

the system, the nested activities and objectives pursued at each point within the system should make sense within the global vision of the systemic reform. The holistic view represents the *complexity* of the systemic reform.

Third, evaluation of systemic reform should offer a *systems view*. Principally, a systems view requires that evaluation examine the totality and complexity of the reform operating as a united force within the system (Banathy, 1995). A systems view suggests that evaluations should ultimately focus not on the degree of success of isolated components of the reform, but rather on the degree of success of the whole reform within the whole system through attention to a shifting balance of linkages, interdependencies, and processes that amplify desired outcomes (Webb, 1997). Furthermore, evaluation that adopts a systems view will not attend to discrete causes and effects, but rather to evidence that the reform effort as a whole contributes to successful solutions to entrenched problems throughout the system (Julian, Jones, & Deyo, 1995). An evaluation with a systems view will provide information and judgments regarding priorities of the reform, sequencing of activities, development and evenness of quality learning throughout the system, connections needed to achieve success, and barriers that might preclude success (CPRE, 1995; Julian, Jones, & Deyo, 1995). The systems view represents the *integrity* of the systemic reform.

Finally, evaluation of systemic reform should serve the understanding and management of systemwide change through the value that it adds to development of the reform. Evaluation might provide warnings about potential challenges and failures; identification of opportunities to link with other efforts; and indications of schools, districts, or components of the reform that should be given extra attention at various times (O'Day & Smith, 1993; Ridgway, 1998). The processes of evaluation ought to promote learning throughout the system through self-reflection against criteria that stakeholders trust (Corcoran, 1997; Goertz, Floden, & O'Day, 1996).

Tracking Change Over Time

Systemic reform is an endeavor that is expected to last many years. Evaluations of systemic reform must take this temporal dimension into account by providing evidence of and supporting judgments about the value of systemic reform over time (Clune, 1993). Although many researchers and evaluators agree that the full impact of systemic reform on any school, district, or state cannot be assessed for many years, evaluation can make critical contributions to tracking change over time from the outset of a systemic reform effort (CPRE, 1995; Goertz, Floden, & O'Day, 1996).

First, evaluations should include needs assessments for the system as a whole and for different parts and functions of the systems; identification of important baseline indicators should follow from needs assessments (I. Weiss, 1997). From needs assessments and baseline indicators, evaluators will gain meaningful ideas regarding "what to look at" in order to track meaningful change.

Second, given the systemic reform's strategies and time lines, evaluations can make conjectures regarding when, where, and to what extent certain changes in the system's infrastructure or its outcomes might be expected (Ridgway, 1998). From these conjectures, evaluators should derive insight regarding "where and when to look" for meaningful changes.

Third, continuous evaluative feedback on implementation and early change will provide important information about opportunities and challenges that managers may use to reorient and reposition the strategic thrusts of the reform. Evaluations should continually determine "what lessons have been learned" that can aid development. Past and current evaluations have revealed that designs of systemic reform evolve considerably over time. Evaluations can make a significant contribution to the ongoing design of systemic reform efforts if they alert key audiences to the greatest needs and the most potentially beneficial opportunities at various times in the life of a systemic reform effort (Heck & Webb, 1996; Ridgway, 1998).

Evaluators who examine change over time should never lose sight of two vital principles underlying systemic reform. First, change itself is not the intent of systemic reform; valuable development toward ambitious learning goals is the aim. To provide evidence regarding how a system has developed with respect to desired effects and impacts will be far more powerful and useful than merely demonstrating that the system has changed (Heck & Webb, 1996; Rowland, 1994). Second, in evaluation of systemic reform assigning blame or praise for past action should remain distantly subordinate to reflection, learning, and guidance for future development. Systemic reform is about growth toward a vision of the future and evaluation primarily ought to track change and inform development toward that vision (Banathy, 1995; Rowland, 1994). Both evaluators and reformers need to have a view of the past, the present and the future in order to understand, first, how the present state of the system represents learning from the past; second, how the present state relates to the idealized future vision of learning; and third, how the systemic reform might make the future vision possible.

Summary

Evaluation of systemic reform can serve a number of important purposes; most evaluations will address several purposes at once. Three critical purposes that evaluation, particularly developmental evaluation, can serve well have been introduced, and two have been highlighted. First, evaluation of systemic reform should aid stakeholders' understanding and management of systemwide change. Systemic reform generally involves multiple, interacting components targeting different functions in numerous local sites. Evaluation can be a valuable tool for managing the developmental challenges of such complex efforts if it consistently represents the totality, complexity, and integrity of systemic reform. Second, evaluation of systemic reform should track change over time. The long-term, progressive nature of systemic reform

demands that reformers and stakeholders understand how the reform develops toward the vision that guides it. Well-designed evaluation can trace the development and adaptation of the systemic reform to changing conditions and ideals. Third, although not directly addressed in this paper, evaluation can help build and test the theory of systemic reform. The theory of systemic reform is the means by which to describe, interpret, and learn from enactments of systemic reform. Evaluation can be a vehicle facilitating healthy development and interplay between theory and action.

References

- Banathy, B. H. (1995). Developing a systems view of education. *Educational Technology*, 35 (3), 53-57.
- Bruckerhoff, C. A. (1997). *Lessons learned in the evaluation of the Statewide Systemic Initiatives*. A paper presented to the Evaluation and Policy Studies Team at the National Institute for Science Education, Madison, WI, March 13-14, 1997.
- Clune, W. H. (1993). The best path to systemic educational policy: Standard/centralized or differentiated/decentralized? *Educational Evaluation and Policy Analysis*, 15 (3), 233-254.
- Consortium for Policy Research in Education. (1995). *Reforming science, mathematics, and technology education: NSF's State Systemic Initiatives*. New Brunswick, NJ: Author.
- Corcoran, T. B. (1997). The role of evaluation in systemic reform. In W. H. Clune, S. B. Millar, S. A. Raizen, N. L. Webb, E. D. Britton, D. C. Bowcock, R. L. Gunter, & R. Mesquita (Eds.), *Research on systemic reform: What have we learned? What do we need to know?* (Synthesis of the Second Annual NISE Forum, Volume 2: Proceedings, Workshop Report No. 4, pp. 64-68). Madison, WI: National Institute for Science Education, University of Wisconsin-Madison.
- Cronbach, L. J., Ambron, S. R., Dornbush, S. M., Hess, R. D., Homik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.
- Goertz, M. E., Floden, R. E., & O'Day, J. (1996). *Systemic reform* (Volume I: Findings and conclusions). Washington, DC: U. S. Department of Education Office of Educational Research and Improvement.
- Heck, D. J., & Webb, N. L. (1996). *Purposes and issues of systemic evaluation in education as reflected in current evaluations and literature*. Unpublished manuscript.
- Jenness, M. & Barley, Z. (1995). Using cluster evaluation in the context of science education reform. *New Directions in Program Evaluation*, 65, 53-69.
- Julian, D. A., Jones, A., & Deyo, D. (1995). Open systems evaluation and the logic model: Program planning and evaluation tools. *Evaluation and Program Planning*, 18 (4), 333-41.
- LeMahieu, P. B. (1997). The role of evaluation in systemic reform. In W. H. Clune, S. B. Millar, S. A. Raizen, N. L. Webb, E. D. Britton, D. C. Bowcock, R. L. Gunter, & R. Mesquita (Eds.), *Research on systemic reform: What have we learned? What do we need to know?* (Synthesis of the Second Annual NISE Forum, Volume 2: Proceedings, Workshop Report No. 4, pp. 69-70). Madison, WI: National Institute for Science Education, University of Wisconsin-Madison.
- National Science Foundation. (1993). *An overview of the National Science Foundation's Urban Systemic Initiatives program*. Unpublished manuscript.
- O'Day, J. A. & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 250-312). San Francisco: Jossey-Bass.
- Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice*, 15 (3), 311-319.
- Ridgway, J. (1998). *The modeling of systems and macro-systemic change: Lessons for evaluation from epidemiology and*

- ecology (Research Monograph No. 14). Madison: University of Wisconsin-Madison, National Institute for Science Education.
- Rowland, G. (1994). Designing and evaluating: Creating futures and appreciating error. *Educational Technology*, 34(1), 10-22.
- Webb, N. L. (1997) The role of evaluation in systemic reform. In W. H. Clune, S. B. Millar, S. A. Raizen, N. L. Webb, E. D. Britton, D. C. Bowcock, R. L. Gunter, & R. Mesquita (Eds.), *Research on systemic reform: What have we learned? What do we need to know?* (Synthesis of the Second Annual NISE Forum, Volume 2: Proceedings, Workshop Report No. 4, pp. 71-73). Madison, WI: University of Wisconsin, National Institute for Science Education.
- Weiss, C. H. (1991). Evaluation research in the political context: Sixteen years and four administrations later. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century* (Nineteenth Yearbook of the National Society for the Study of Education, Part II, pp. 211-231). Chicago: University of Chicago Press.
- Weiss, I. R. (1997). The role of evaluation in systemic reform. In W. H. Clune, S. B. Millar, S. A. Raizen, N. L. Webb, E. D. Britton, D. C. Bowcock, R. L. Gunter, & R. Mesquita (Eds.), *Research on systemic reform: What have we learned? What do we need to know?* (Synthesis of the Second Annual NISE Forum, Volume 2: Proceedings, Workshop Report No. 4, pp. 73-75). Madison, WI: University of Wisconsin, National Institute for Science Education.

OVERHEADS USED

Purposes and Vision for Evaluation of Systemic Reform

Daniel J. Heck

University of Illinois at Urbana-Champaign
NISE Forum, 1999

1

Evaluation and Systemic Reform

Evaluation "tends to *ignore the social and institutional structures* within which the problems ... are generated and sustained."

"Most of the political implications of evaluation have an *establishment orientation*. They accept--and bolster--the status quo." (Weiss, 1991)

2

Developmental Evaluation

"Evaluation processes and activities that *support program ... development*. The evaluator is part of a team whose members collaborate to *conceptualize, design, and test* new approaches in the *long-term, ongoing process of continuous improvement, adaptation, and intentional change*." (Patton, 1994)

3

Purposes for Evaluation of Systemic Reform

- To aid understanding and management of change throughout the system
- To track the nature and extent of change over time
- To build and test a theory of systemic reform

4

Understanding and Managing Systemwide Change

Perspectives

- Whole System View--System, Reform, and their Intersections
- Holistic View--How each activity or objective relates to primary goals and overall vision
- Systems View--Totality and complexity of reform as a united force

5

Understanding and Managing Systemwide Change

Quantitative Techniques

- Hierarchical Linear Modeling
- Structural Equation Modeling

Qualitative Techniques

- Nested Case Studies
- Case-Ordered Displays and Analyses

6

Tracking Change Over Time

Perspectives

- Baseline Indicators--What should we look at?
- Conjectures, Projections--When and where do we look?
- Feedback on Implementation--How do things look now?
- Ongoing Design--What we want things to look like in the future?

7a

Tracking Change Over Time

Quantitative Techniques

- Time Series Analyses
- Repeated Measures Designs
- Hierarchical Linear Modeling

Qualitative Techniques

- Longitudinal Case Studies
- Time-Ordered Displays and Analyses

7b

Summary

- Developmental evaluation is an approach that is well-matched to important purposes for evaluation of systemic reform
- Evaluation should aid in management, leadership, and understanding of systemic reform

8

TRACKING THE THEORY OF CHANGE: A MOVING TARGET

Evaluation of Systemic Reform in Mathematics and Science

Zoe A. Barley
Western Michigan University

Background

Education's move toward systemic reform arose from prior failed efforts to improve educational outcomes for students by focusing reform separately on changing teaching, instructional materials, or curricula. While each of these might have changed initially, the larger school context eventually defeated realization of the desired improvement in student outcomes. In some cases, change itself became impossible given the barriers, e.g. policies, procedures, or resource scarcity presented by the school context. In other cases gains in one area were offset by losses in another. Educators and researchers came to see these barriers, not as isolated issues, but as part of a system of education. All of the pieces—including roles and relationships of the persons involved, policies and procedures, resources and capacities—needed to be understood as they interrelated to support or defeat reform. Comprehensive systemic reform underscores the necessity for reformers to consider all aspects and influences that finally determine how students develop the knowledges and skills desired. This extends the reform to include parents and community and to other levels of influence such as higher education, state educational policy makers and federal policies and programs. Comprehensive systemic reform is now understood to be the essential strategy for educational reform.

The National Science Foundation (NSF) was an early and strong supporter of systemic change initiatives. Through its Systemic Initiative (SI) programs, states (SSIs), rural (RSIs), and urban (USIs) areas were funded to conduct systemic reform with an emphasis on improvement in mathematics and science. Each of these initiatives was required to have an external evaluator and for the statewide initiatives, SRI International was contracted to

conduct a national evaluation. Many of the external evaluators attended biannual conferences held to support networking among Principle Investigators and Project Directors of the state initiatives. These furthered the dialog among evaluators about issues in the evaluation of systemic reform. Eventually evaluation issues led to an SSI evaluation workshop (February 1994) on using logic models (Rog, 1994) to link evaluation and key SSI strategies.

Logic models are an off-shoot of Chen's (1990) work on theory-driven evaluation. Theory-driven evaluations seek to elucidate the program theory, "the set of interrelated assumptions, principles, and/or propositions to explain or guide social actions" (Chen, 1990, p. 40) that the program designers have in mind, consciously or unconsciously, as they develop and implement the program being evaluated. Program theory became more complex with the move to systemic reform, encompassing the entire system relevant to the desired outcomes including context, presumed causal factors, mediating factors, and the interventions or program activities themselves. Schon (1997), in his work on program theory, noted the importance of paying attention not only to the espoused theory, the originally developed understanding, but also to the theory of action (or theory-in-use), which emerges as the program or reform is implemented.¹

Some SSIs developed logic models for their state systemic reforms as a result of the workshop and the evaluators to a greater or lesser degree used these in shaping the SSI evaluations. This paper reports on one evaluation team's experience with the use of theory-driven evaluations for several systemic

¹ Schon distinguishes a third theory, the "design theory," which emerges as the espoused theory gets concretized in budgets and planned actions.

reform initiatives including two SSIs, an OERI funded advanced technology grant, and a privately funded urban systemic reform.

Critical Factors in the Efficacy of Systemic Reform Initiatives

The first task of a theory-driven evaluation is to give form and specification to the theory, drawing upon program documentation and directed conversations with the program directors. A theory-driven evaluator needs to have a good understanding of best thinking in the program area in order to describe the program theory and to identify gaps in the logic or misperceptions in what will accomplish the outcomes. For systemic reforms, given their complexity, this need is even more important. The NSF Office of Systemic Reform program staff developed definitions for a set of eight elements of systemic reform and six "drivers" that NSF perceived to be essential in moving reform forward. Other documents that emerged to define systemic reform for SSIs included "A Continuum of Systemic Reform" developed by Beverly Anderson (Education Commission of the States, 1992) and SRI's concept model for the SSI national evaluation (SRI, 1992). These additional documents, as well as emerging research about successful systemic reform initiatives were useful in identifying gaps or misperceptions in the program theory of the reform.

Chen (1990) makes the distinction between "normative theory," what the structure of the program should be (prescriptive), and "causative theory," what the underlying causal mechanisms actually are (descriptive). The normative theory is what the evaluator finds in examining the program documents and talking with program directors. It usually has come from unexamined premises or prior experience. The causative theory is empirically based and comes from the relevant literature. The evaluator explicates both theories in order to design a theory-driven evaluation. The normative theory assesses the consistency of the actual program activities in relation to the intended intervention and shapes the

evaluation of the implementation. The causative theory assesses both the impact of the program and how the impact was generated and shapes the summative evaluation.

Three distinctive features of systemic reform influence the design of the evaluation. Inevitably inherent within the understanding of the program theory is a set of values. Minimally, the intended outcomes are *valued* for those who will participate in the program. Systemic reform also includes value at a systems level, the value of continuous improvement, a self-renewing process in which the system makes corrections in its strategies and processes to enhance the attainment of desired outcomes. Such a process entails the collection, analysis, and interpretation of data as a part of system functioning. While the external evaluation could operate entirely independent from the internal evaluative process, more typically the evaluators have operated in a supportive role, providing technical assistance for the internal evaluation process while gaining data useful for the external evaluation. Close collaboration of the two efforts reduces the data burden on the elements of the system and maximizes the information available to program directors and funders. This first feature, the collaborative role, is then reflected in the evaluation design. A second typical characteristic of systemic reforms is the press for stakeholder involvement at all stages. Stakeholder involvement in the external evaluation is best served if a representative group of stakeholders is involved beginning at the design stages. Finally, systemic reform is a long-range process. Evaluators must identify or develop intermediate benchmarks as a means to assess whether the reform is progressing prior to expected changes in ultimate outcomes. This third feature thus adds another dimension in evaluation design.

Developing a Theory-Driven Systemic Reform Evaluation

As the evaluator works to make explicit the theory or logic of the systemic reform,

Schon (1997) suggests a series of pertinent questions: Is the design theory congruent with the espoused theory? (Do we design what we espouse?); Is the theory-in-use congruent with the design theory? (Do we enact what we have designed?); Are the theories internally consistent?; and, Is a given theory of action effective in the sense that its strategy yields the desired outcomes? One way to present the design theory is what is known as a logic model, a conceptual representation of the relationships among the relevant inputs, intervening factors, intermediate benchmarks, and interventions leading to the outcomes. The logic model is developed based on the data at hand. In the early stages, this data would come from planning documents, early implementation pieces and interviews with program staff. Presuming the evaluator is on board as the detailed planning is conducted, the logic model can serve not only to assist the program planners with identifying program design problems but for the evaluator it may serve as an evaluability assessment (Wholey, 1979), an analysis of whether the reform is sound enough in design to warrant an evaluation. The next steps for the evaluator are to enter into a dialogue with program staff about the logic-or absence of logic-in the planned work, to develop the collaborative relationship with internal evaluators, and to engage other stakeholders in final aspects of the evaluation design.

In one systemic reform, we found that the program staff, who were dedicated program activists, lacked interest in the logic model approach and could not grasp the importance of assessing whether the detailed plans would actually realize the outcomes their espoused theory promised. For the first few years of this initiative, the implemented program was actually a multiplicity of separate programs with little or no "systemicness" about the work. The desired outcome of a reformed system did not occur despite the realization of many useful outcomes for individual teachers. In another systemic reform initiative, the program staff worked hard at developing the logic model. In the process they realized that a number of related agencies would need to be brought on board if the reform were to

succeed. Building this broader base of involvement became a key strategy and the interconnections an important intermediate outcome.

In both cases cited above, the evaluation was designed using a combination of the theory of action of the program implementers and the causative theory-the best thinking available about systemic reform initiatives. The logic model, even when not seen as useful, served as a means to negotiate evaluation emphases for the evaluation design. For the first case while system change was tracked, the emphasis was on the particular results of various program efforts. For the second case, much more emphasis was placed on changes in the **system**—policies, procedures, extant programs—and in the connections among the key system elements. Interestingly, funders were more interested in specific program outcomes than in system changes despite espousing systemic reform.

Implementing a Theory-Driven Evaluation

Because the development of logic models came after the initiation of the evaluations for the SSIs, the models represented the theory after it had been significantly modified as a result of the management team better understanding what systemic reform entailed. In another case, the evaluation of an advanced technology grant, we were able to initiate the theory-driven approach including a logic model as the program began. The program plan entailed installing equipment centrally and in classrooms and setting up electronic connectivity for all teachers across 5 districts in a single county. The desired outcomes included student achievement improvement, teacher retention, and teacher and student use of electronic support for teaching and learning. The logic model revealed gaps between the planned work and the desired outcomes. When this was revealed the consortium sought GOALS 2000 money to institute professional development for the teachers in the use of technology and to establish a common curriculum to foster connections among teachers. While gaps

remained in the "logic" of the work, the additions suggested the program warranted evaluation (evaluability assessment) and the relationship forged with the program implementers created a continuous learning mode allowing data collected by the evaluation to continue to inform program decision-making.

For the privately funded urban reform, a logic model was developed as part of the response to the RFP for the evaluation. During the six-month evaluation-planning period the model was further refined with input from the community-based Evaluation Committee. For the reform to be successful, a delicate balance among business, strong locally focussed foundations, community activists, and the school district had to be worked through. The presenting issues were always couched in who had authority or power and what access non-district groups and persons would have to district decision making. Yet the goals of success for all students were commonly held by all. The original model had to be rethought to better represent untested theories about a community role in a large hierarchical urban district as a necessary precursor to radical rethinking of teaching and learning. As evaluators rethought the logic model, a reallocation of evaluation resources was necessary. More effort was expended than first planned in monitoring district policies and personnel changes and in observing the emerging relationship between key community leadership and the central administration of the district. The establishment and institutionalization of a set of common success indicators to be measured by the district but developed collaboratively and monitored by the evaluators was one important coming together. Evaluation resources were also redirected to support this collaborative work, a fairly labor intensive effort in engaging sometimes hostile community members with isolated and defensive district assessment personnel to come to consensus on the definition and measurement of the indicators. In this case, ongoing reflection on the theory of action held by the various parties enabled the evaluation

team to better focus its work and interpret the data collected.

Some Concluding Thoughts

Thus, a theory-driven approach to systemic reform evaluation may ultimately be of greatest use to evaluators in shaping reflection and focusing their work. It does, however, especially in conjunction with a logic model, offer a better way for reformers to graphically understand not only the results they seek in relation to the strategies they undertake, but also the multiplicity of factors in the larger system in which they operate.

As a part of the evaluation design, data collected on intermediate benchmarks provide early checks on whether the reform is on track. A collaborative relationship with system reformers influences the role evaluators play, for example adding technical assistance and consensus building, but it also can result in better and more efficient data collection as responsibility is shared and results mutually beneficial. Evaluators also gain tools for their own reflection from a theory-driven approach including an early consideration of whether a reform warrants an evaluation effort.

References

- Anderson, B. (1992). *The state education system-A continuum of systemic change*. Denver, CO: Education Commission of the States.
- Chen, H. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage Publications.
- Rog, D. (1994). *Using logic models to link evaluation and key SSI strategies*. Unpublished.
- Schon, D. (1997). *Notes for a theory-of-action approach to evaluation*. Unpublished draft.
- Shields, P. M., & Zucker, A. A. (1992). *Study of NSF's Statewide Systemic Initiatives (SSI) Program*. Menlo Park, CA: SRI International.
- Wholey, J. S. (1979). *Evaluation: Promise and performance*. Washington, DC: The Urban Institute.

Requirements for the 21st Century

- An interdisciplinary environment that challenges the way we organize classrooms, subjects and knowledge in schools and colleges
- A curriculum that stresses lifelong **skills** such as learning how to learn instead of rote teaching
- Teachers who take on different roles • not only lecturing, but also coaching, role playing, facilitating
- An environment that encourages students to take more responsibility for their learning, becoming active participants instead of passive recipients of information
- Improved forms of assessment, including portfolios, exhibitions and demonstrations
- Creative use of time and space
- More individualized instruction, i.e., methods that respond to students' individual learning styles
- More diverse ways of organizing and presenting information
- More decision-making autonomy given to those persons closest to the problems

From Introduction to Systemic Education Reform • EDC

REFLECTIONS ON SYSTEMIC REFORM EVALUATION: NISE CONFERENCE

Issues in Evaluating Systemic Initiatives:

If it is intended that the evaluation itself model/embody a systemic approach, the following ensue:

- the list of constraints is not just doubled but squared - given the confounding effects.
- the complexity of the evaluation constantly challenges the ability to focus
- expectations of stakeholders, about the criteria for evaluating, the role of evaluators, their roles given shared decision-making, etc are difficult to sort out and meet
- the political aspects of the evaluation are not only the context but also influence the release and use of findings
- new roles are required of evaluators who already lack experience in the task at hand

Problems in Evaluating Systemic Initiatives:

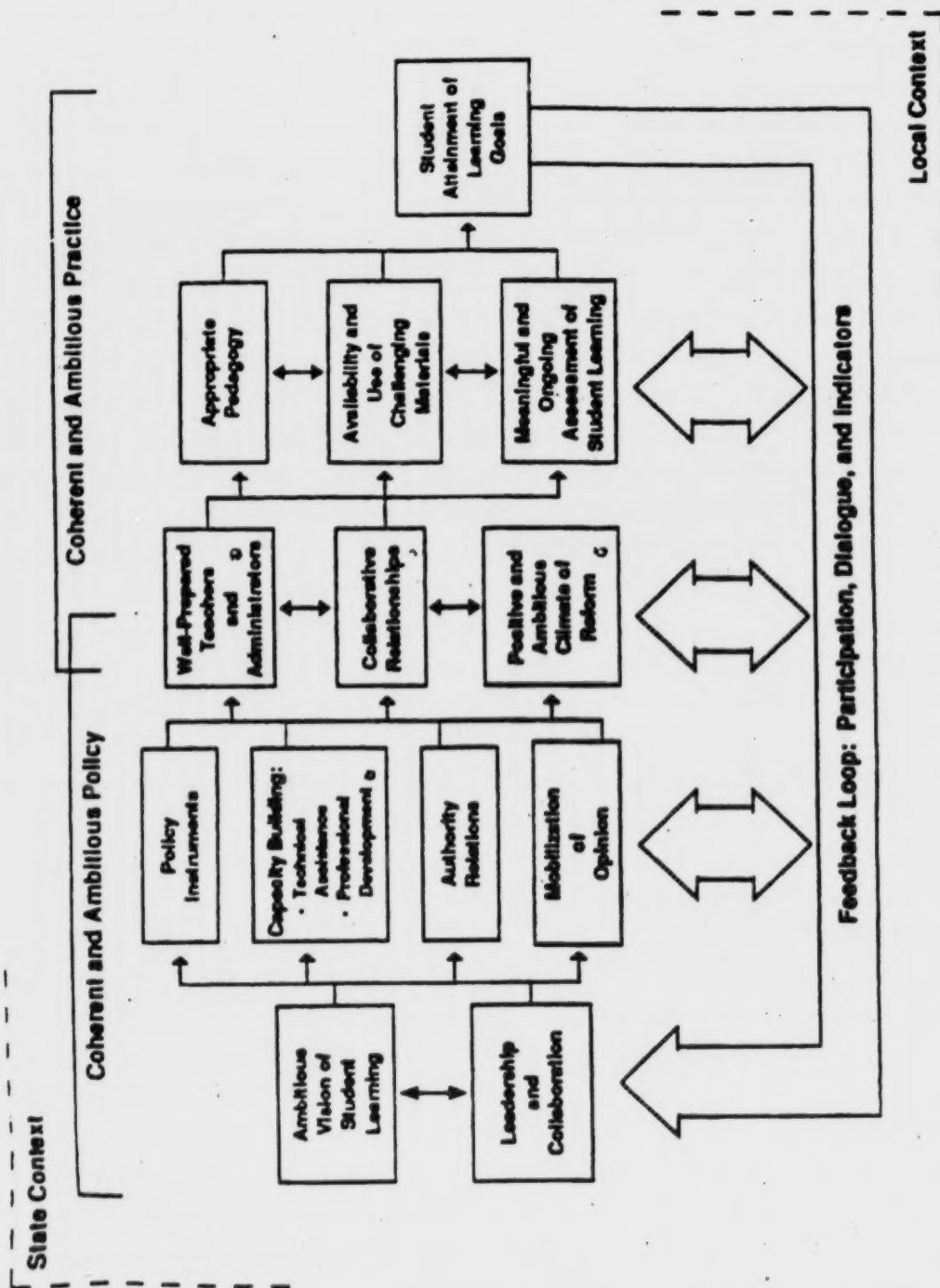
- getting program implementers who are activists to use evaluation findings in making decisions about revising program directions
- teaching/instilling a mindset and methods for using evaluation data from a non-teaching platform
- operating within a "learning community" as an evaluator yet member of the community
- if evaluation recommendations are adopted, are evaluating one's own program directions
- allocation of resources in a complex, messy, emergent design evaluation is a constant issue

ESSENTIAL ELEMENTS OF A PROGRAM LOGIC MODEL

<u>The Beginnings:</u>	<u>The Planned Work:</u>	<u>The Intended Results:</u>
Assumptions	Strategies	Short Term Outcomes
Problem Statement	Activities	Long-term Outcomes
Needs Identified	Programs/Events	Goals

Panel I:
NISE- 1999

Tracking the Theory of Change: A Moving Target
Zoe A. Barley, SAMPI - Western Michigan University

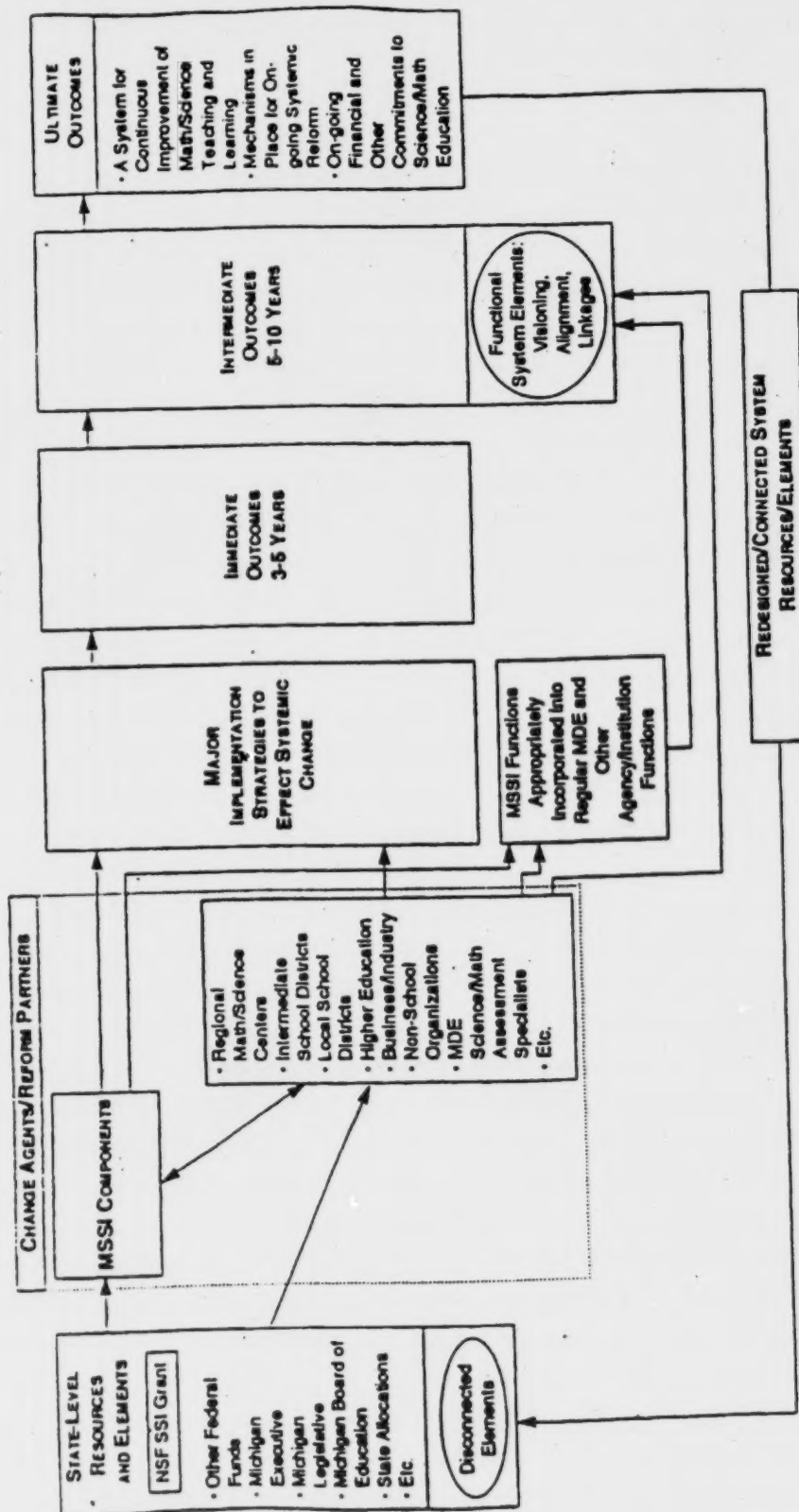


Draft - 1 - 6

FIGURE 1 A FRAMEWORK FOR EVALUATING SYSTEMIC REFORM

**MICHIGAN STATEWIDE SYSTEMIC INITIATIVE
PROGRAM LOGIC MODEL**
Summary Version*

Draft 5/94



Prepared by MSSI Evaluation Team
Science and Mathematics Program Improvement (SAMPPI)
Western Michigan University
Kalamazoo, MI 49008
Phone 616 387-3791

*A detailed version of this model is available.

QUESTIONS TO GUIDE THE EXPLICATION OF LOGIC
(Schon, 1997)

Is the design theory congruent with the espoused theory?

Is the theory-in-use congruent with the design theory?

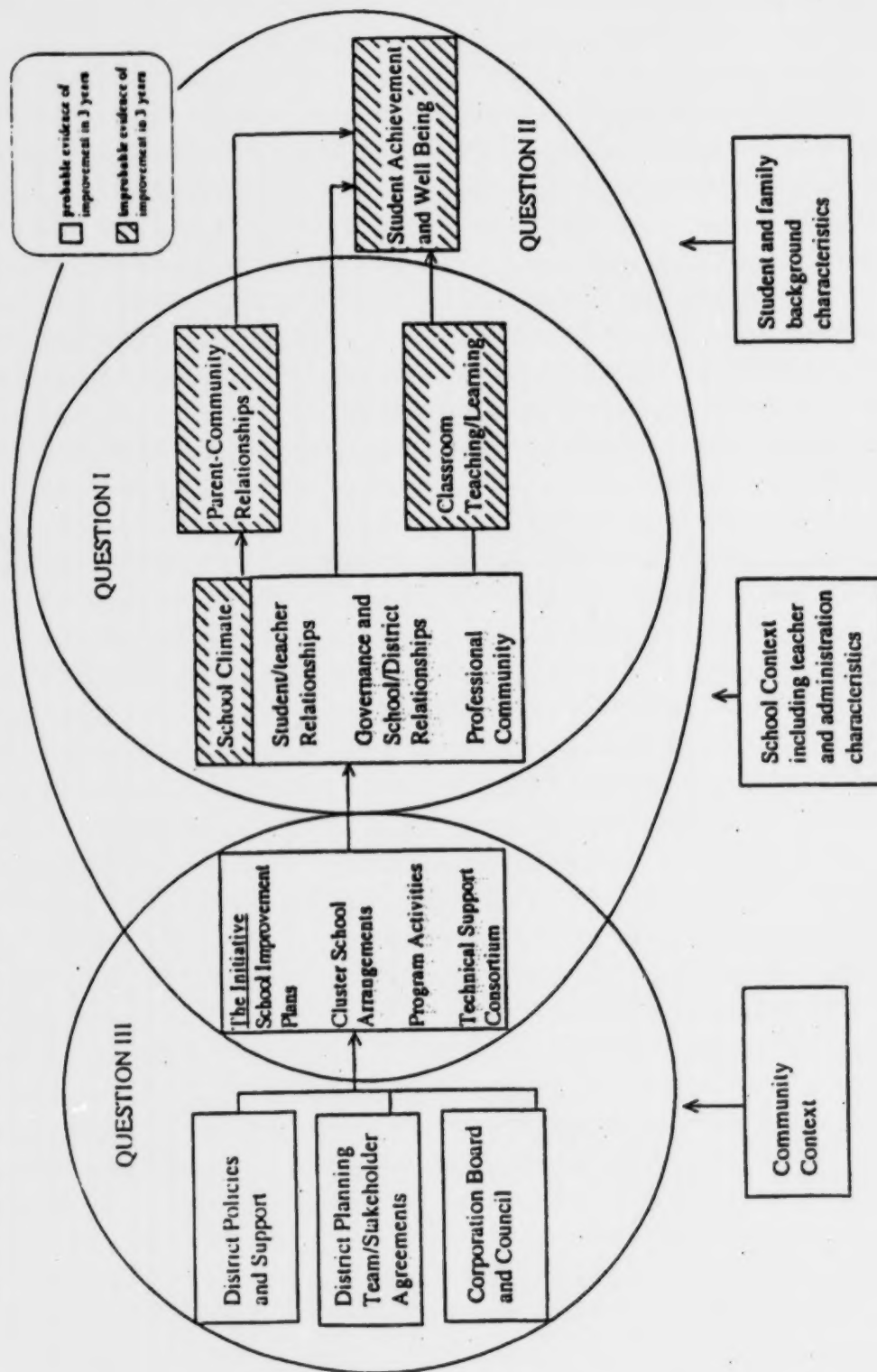
Are the theories internally consistent?

Is a given theory of action effective in the sense that the strategy yields the desired outcome?

Panel L:
NISE- 1999

Tracking the Theory of Change: A Moving Target
Zoe A. Barley, SAMPI - Western Michigan University

Exhibit 2. Concept Map and Major Evaluation Questions: Schools of the 21st Century



REVISED 10/97

EVALUATING SYSTEMIC REFORM: A COMPLEX ENDEAVOR

Iris R. Weiss
Horizon Research, Inc.

At first glance, evaluation of systemic reform efforts is a lot like evaluation of any reform effort, where information is collected, analyzed, and interpreted in order to: (1) improve the project and/or (2) assess its impact. But the fact that systemic reform efforts are charged with aligning the many components of the education system that make the evaluation efforts substantially more challenging.

Leaders of mathematics and science education reform efforts, both "systemic" and "non-systemic," typically begin by laying out the needs they are addressing. They use their understanding of the system, in concert with their knowledge of what works best in a particular context, to design a set of interventions. Evaluation can begin very early in the reform process, with evaluators using their knowledge from research and prior experience to critique the design of the initiative and suggest areas in need of refinement, but relatively few projects use evaluators in this role. More typically, once the reform begins, evaluators monitor the quality of the implementation and seek evidence of its impact. By sharing the results of the evaluation with key project stakeholders, evaluators hope to contribute to continued improvement of the reform design and implementation. By documenting areas of impact (and lack of impact), evaluators hope to inform future policy and program decisions.

In reforms of an individual component of the education system, evaluators can use time-honored evaluation strategies with reasonable confidence. In contrast, in evaluating systemic reform efforts, evaluators often feel like we are making it up as we go along. The following sections describe the (relative) ease of evaluating a traditional intervention, and the increased complexity involved in

evaluating multi-faceted systemic reform initiatives.

Evaluating "Traditional" Reform Efforts

For many years, mathematics and science education reform efforts focused on individual components of the system. For example, summer institutes were offered at colleges and universities to help in-service teachers deepen their content knowledge. Evaluations of such efforts looked at the quality of the institutes in relation to that goal (e.g., to determine if the content was important for teachers to understand and was presented in a way that was accessible to those teachers). They found out from teachers whether the institutes had impacted their feelings of preparedness. On occasion, they might use pre- and posttests to determine more objectively if teacher knowledge had increased. The evaluators might have used their experience with similar programs to critique the design ahead of time, and they would likely provide formative feedback as the project unfolded. However, since interventions typically were not attempting (or likely) to change the educational system beyond impacting teacher preparedness, there was no reason for evaluation efforts to focus on the larger system, or for the evaluators to provide feedback in regard to changing other parts of that system.

Evaluating "Simple" Systemic Reform Efforts

By definition, systemic reform efforts address multiple components of the education system. As a result, the evaluator's job expands beyond tracking the implementation and impact of the specific project activities to looking at other elements of the system that

may affect the extent to which the project achieves its goals. Since everything in a system is connected to everything else in that system, it is often unclear where to draw the line in deciding what is and is not to be addressed in the evaluation, especially when projects funded as systemic are only minimally so.

What *is* clear is that even "simple" systemic reform efforts introduce complexities for evaluation. Consider the case of the National Science Foundation's (NSF) Local Systemic Change through Teacher Enhancement (LSC) initiative, which emphasizes professional development around exemplary instructional materials. Assume that in a particular district the key needs have been identified as:

- Many elementary teachers lack science content knowledge, and
- Elementary teachers are not prepared to use the instructional materials the district has chosen.

The literature on systemic reform led the project staff to: (1) design professional development programs in which the elementary teachers could learn content and pedagogy in the context of the designated instructional materials; (2) work with principals to ensure that they do not derail the reform efforts; and, (3) provide support for teachers as they attempt to implement the new instructional materials in their classes.

Obviously, the fact that the project activities address more components of the system than simply teacher content knowledge increases the scale and complexity of the evaluation, which now must focus on several areas. Less obvious is the fact that the evaluation must now also consider parts of the system *not* directly addressed in the project plan. For example, in critiquing the project design, evaluators might suggest the need for a materials management center, noting that other elementary science projects they have evaluated have floundered when teachers had to deal with re-supplying consumables. Similarly, in looking at impact, the evaluation would need to consider issues of

sustainability, i.e., whether there was a system in place for the district to continue the professional development after the grant. Consequently, not only will the evaluation of a systemic reform effort require resources beyond those needed to evaluate a similar size traditional reform effort, but it will also require evaluators with a broader and deeper understanding of educational systems.

Evaluating "Complex" Systemic Reform Efforts

Many systemic reform efforts are considerably more complex than the LSC example, and the ensuing challenges for evaluation increase correspondingly. Let's look at how increasing the complexity of a systemic reform effort complicates the evaluation, both in the "design critique" stage and in the evaluation of the project's implementation and impact.

Design Critique

The LSC solicitation specified that the reform was to emphasize professional development. At the same time, projects were asked to situate that professional development in a systemic context so that other aspects of the system did not negate the impact of the professional development. (For example, if the district assessments were not consistent with the content and approach of the new instructional materials, teachers would be less likely to change their instruction.) Once the subject and grade range for the intervention was designated, and the goals identified, the evaluation would be targeted to that subject, that grade range, and those specific goals. The evaluators did not need to focus on whether it would have been better to spend project resources on pre-service education, revising the high school science or mathematics curriculum framework, or any of a myriad of interventions that were outside the scope of this initiative.

In a more complex systemic reform effort, leaders' attempts to "understand the system" could well generate a lengthy list of **needs**—for professional development at every grade

range; for an articulated K-12 science, mathematics, and technology curriculum; for improved instructional materials; for assessments aligned with reform; for administrator support; for higher expectations for all students; for community support; for replacing antiquated laboratories; for improving pre-service preparation.

If systemic reform theory were well-developed, project staff would have some direction for deciding how much priority to give each need and in what sequence, and evaluators would have a sound basis for critiquing the project design. But systemic reform theory is exceedingly thin, specifying overall goals, but providing little guidance on how to go about meeting those goals. There is a bewildering array of options for intervention, and often as many opinions about the most effective strategies as there are stakeholders. In a simpler systemic reform effort, the evaluator's critique would likely help the project improve its design. In a more complex endeavor, the evaluator's voice may simply add to the confusion.

Implementation Evaluation

Every systemic reform initiative eventually settles on a course of action, a subset of the seemingly infinite number of activities which have the potential to address the system's needs. In most cases, however, that subset is still much more than the evaluators can possibly monitor within the time and resources available. Typically, the resources devoted to evaluation in systemic reform efforts would be more appropriate for investigating the quality and impact of two or three components, not the dozen or so that are generally included.

In the best circumstances, project staff help decide where to target evaluation resources both by being clear in communicating the project strategy and in specifying the programmatic decisions that will need to be made. In the more typical case, project staff want an in-depth look at everything, or the various stakeholders are interested in different parts of the initiative or different parts of the system. The process of

reaching consensus requires extended, sometimes seemingly endless negotiations. Eventually data collection begins, whether according to an agreed-upon evaluation design or more haphazardly, simply because the clock is running and evaluators need to have something to report. At this stage, it is possible to pretend that this is a typical evaluation, proceeding to review project documents, observe project events, interview participants, talk with key stakeholders in the system, etc.

Typically, the plot thickens when it is time to provide formative evaluation feedback. In a simpler project it might be appropriate to report results only to the Principal Investigator (PI), but the collaborative approach inherent in complex systemic reform efforts suggests the need to communicate with a larger group. In fact, even if the systemic reform has a single dominant leader, it is a good idea to communicate evaluation findings more widely. Intentionally or otherwise, the PI may filter the information, put a "spin" on it, or use the results in some other way that seems counter to the best interests of the initiative. To avoid this problem, we have learned to provide feedback in writing simultaneously to the project's entire management team, typically 3-10 people, leaving it up to them to decide who else should get the report and when.

In any evaluation, but especially in complex systemic reform efforts, there is an additional problem in finding the appropriate point in the balance between simply report findings versus making recommendations for a major redesign to increase the likelihood of impact. At one extreme, the project is deprived of the insights of skilled, experienced people who understand the project goals and context deeply and well. At the other extreme, those same skilled, experienced people could be perceived as taking over the project, and in turn, evaluating themselves!

A related challenge is presenting information in a way that will help the project move forward. In the ideal, project staff would have both the capacity and the will to

make use of evaluation feedback to improve the project design and implementation. The reality is, unfortunately, very far from that ideal, especially in complex systemic reforms.

We have found a number of reasons why projects are unable to make mid-course corrections, even in the face of compelling evaluation results. In developing the initial reform plan, project leaders often had to negotiate with diverse stakeholders, and they may be concerned that any changes will jeopardize the sometimes fragile coalition that was established at that time. Alternatively, project staff may know how to do what they proposed initially, e.g., high-quality professional development, but not know how to go about whatever it is the evaluation results suggest they do instead, especially if the recommendations involve efforts in the policy arena. Finally, the turf issues that are present in any initiative seem to increase exponentially with the number of players; sometimes formative evaluation feedback in a large systemic initiative becomes just another round of ammunition for the political battles.

Impact Evaluation

Funders have their own constraints, including the need to provide evidence of program effectiveness to Congress or other policymaking groups. Unfortunately, this need often translates into pressure for the initiative to seek evidence of impact when the reform efforts are just beginning to be implemented. At some point, typically long before evaluators think it is reasonable to do so, the evaluation will begin to focus on evidence that the initiative has had its intended impacts on teachers and students. Again, there is likely to be far more to look at than is feasible with the available resources. The problem is complicated greatly by the lack of appropriate outcome measures, a situation that is even more problematic for systemic initiatives than for traditional reform efforts because of the need to demonstrate impact in order to justify the large expenditures.

One difficulty is that systemic reform includes alignment of policy in support of the

reform vision, but in most cases "alignment" has not yet been defined in measurable terms. Another difficulty, even in areas where there are existing instruments, is the scarcity of measures that are simultaneously valid, reliable, and feasible on a large scale. Surveys and multiple-choice tests are open to criticism on validity grounds; classroom observations and performance assessments are open to criticism on reliability grounds, and so on. Finally, there are often problems in study design that threaten the credibility of the results. Unlike small-scale research projects, major systemic reform efforts rarely use random assignment of teachers and students to treatment groups, and appropriate comparison groups are difficult to find.²

Because of these and other complexities, some researchers have suggested that the question of impact on student achievement be addressed through carefully controlled research efforts rather than as a part of the evaluation of professional development interventions.³ The reasoning is that if it can be demonstrated that students learn more when teachers do more X and Y, then evaluation of a particular reform effort could determine if teachers are in fact doing more X and Y, and leave it at that. Politics aside, that advice might be heeded as a more efficient

² Using "matched" districts/teachers/students might work if you chose the "right" matching variables, but the primitive state of systemic reform theory does not inspire confidence in this regard; it is entirely too likely that some unmeasured aspect of the context will make the two groups non-comparable. Choosing as yet "untreated" teachers/students in the intervention districts and schools helps avoid that problem, but introduces the possibility that these groups were influenced by policy reforms associated with the initiative.

³ See, for example, George Hein, "The Logic of Program Evaluation: What Should We Evaluate in Teacher Enhancement Projects?" in *Reflecting on Our Work: NSF Teacher Enhancement in K-6 Mathematics* (S. N. Friel & G. W. Bright, Editors). Lanham, MD: University Press of America, Inc., 1997.

use of evaluation resources. But the final complexity of evaluating systemic reform is that, as in the reforms themselves, there is

virtually no chance that politics will be set aside for very long.

Introduction to Breakout Session I Question Summary

Each panel was followed by a Breakout Session. Participants were assigned to small groups of ten to twelve, led by a facilitator, in a discussion of three questions and other issues raised by the presenters. Each set of three questions was developed by the organizers of the Forum. At the beginning of the Breakout Session, participants were asked to write their responses to each of the three questions on index cards. The comments that the participants wrote were used to begin the small group discussions. The index cards were given to two people, who provided a synthesis of the conference; comments on the index cards were incorporated into their comments. Responses to the first question are summarized here to provide examples of participants' comments.

Breakout Session I: Defining the Problems of Evaluating Systemic Reform

Participants' Comments:

Q: What are the main issues that need to be considered in evaluating systemic reform?

Participants raised some important points about what needs to be considered in evaluating systemic reform. Although a number of issues were identified, a few were repeatedly cited by a number of the 175 participants who responded. These are listed in order, from issues of greatest concern to those less frequently raised.

1. *Clear definition of system, systemic reform, and relevant components.* More than 20% of the participants noted the importance for being clear about what is being evaluated. This requires defining what system is being reformed, including how its boundaries are defined, and determining what is within the system and what should be considered outside the system. Several of the participants thought it important to clearly identify the system components, what is meant by components, and what the interconnections among components are. One participant noted the need for a common framework that can be used to analyze different components, including curriculum and policy. Others emphasized the need to specify clearly what systemic reform is and what the parameters are for a system that is acting systemically (e.g., How much coherence is enough? When is a curriculum standards-based? When is a system serving all students equitably?) A few extended this thought by asking about how to conceptualize systemic evaluation as a complex system, in and of itself, that is in turn embedded in a complex system.

2. *Student achievement.* More than 10% of the participants identified the measurement of student achievement and of systemic reform's impact on student achievement as an important issue. A typical comment was "Bottom line, students-how does [systemic reform] work?"

3. *Ways to work with dynamic and complex systems.* More than 10% of the participants raised the issue of studying education systems that are dynamic, change over time, and are complex. Others raised questions about managing the scale of a large system and selecting the most appropriate operational variables to evaluate the entire system. One participant indicated that the evaluation design for systemic reform has to be responsive to the dynamic nature of the system.

4. *Means for determining attribution and cause and effect.* About 5% of the participants noted that attribution and causality were important issues. A few questioned whether it was necessary to judge attribution. One participant indicated the possibility of assessing only a fractional part of the impact by analyzing the variety of connections between inputs and outputs. Another participant felt that attributing effects to an initiative would require studying the cognitive processes of students and teachers in the learning process.

5. *Responding to and identifying audiences and stakeholders.* More than 5% of the participants made some reference to the relationship of the evaluation and its findings to appropriate audiences. Participants sought clarification on what would be meaningful to different audiences (e.g., initiative personnel, funders, and policy makers), how should results and data be interpreted, and how should feedback of results be varied. A few participants questioned how to evaluate the "buy-in" by stakeholders and the commitment of all constituencies to achieving the desired

outcomes. One participant raised the question about how stakeholders should be viewed and whether they should be considered participants.

6. *Other questions raised by more than one of the participants were:* What are the critical indicators of success that

should be measured? How do we convey the importance of creating logic models and reconciling the research design with the theory/logic of systemic change? How can reform be evaluated fully when there is a misalignment between the existing assessments and the goals and objectives of reform?

Panel II: Models and Approaches to Evaluation of Systemic Reform

Panel Papers and Authors:

Critical Elements of an Evaluation of Systemic Reform

Patrick Shields, Andrew A. Zucker, and Nancy E. Adelman, SRI International

Evaluators' Roles: Walking the Line Between Judge and Consultant

Jeanne Rose Century, Education Development Center

Assessing Student Outcomes

Norma Davila, University of Puerto Rico

Understanding the Value of NSF's Investments in Systemic Reform

Mark St. John, Inverness Research

Discussion Summary and Commentary: Models and Approaches to Evaluation of Systemic Reform

Norman L. Webb

Panel II presented considerations and issues related to models and approaches for evaluating systemic reform. The four speakers discussed critical elements of an evaluation of systemic reform, the evaluator's role, assessing student outcomes, and understanding the value of NSF's investments in systemic reform.

Patrick Shields, a lead researcher of the SSI program evaluation conducted by SRI International, presented the conceptual model the evaluation team developed to specify the key components of an educational system that need to be reformed in concert. SRI based its model on the conceptualization of systemic reform as specified by Smith and O'Day (1991) and others. In the shape of a pyramid, with student outcomes at the apex and standards and institutional collaboration and leadership at the foundation, the model is deceptively simple. Based on clear standards for what students should know and be able to do, and with the support of the key leadership, states and districts must align policy, build capacity to provide schools and teachers with needed human and material support, restructure incentive systems, and build professional and public support for the reform agenda. These actions in turn are meant to

provide the support needed to help increase teachers' capacity to implement the reform vision with access to appropriate materials, within schools organized to support their efforts, and with the support of parents and community. Such synergy will produce reformed classrooms and increased student learning.

The evaluation used quantitative data gathered annually from SSI principal investigators, repeated site visits to each SSI, and reanalysis of data sets gathered by many of the SSIs. Shields noted what proved most useful from their approach to the evaluation. Their evolving model of systemic reform was helpful in guiding their inquiry and in facilitating cross-case comparisons. Twelve in-depth case studies of total state systems effectively provided detailed descriptions and analyses of the progress of the individual SSIs. The evaluation team identified eight specific state strategies that aided in the cross-site analysis and in assessing strengths and weaknesses of sites' focusing on specific components. Shields also noted what they attempted that worked less well. An attempt to develop a common survey to compare classroom data was found to be too difficult. This forced the SRI team to rely on case study and state-selected evaluators' data, which varied in quality, to decipher classroom impact and to assess student learning. Reducing the complex stories of the 26 SSIs into a concise summary, as requested by NSF, and ranking the process of individual states

was found to be daunting. In the end, a report was produced without state-by-state rankings. It proved impossible to identify a small number of models of systemic reform. Instead, the SRI evaluators deferred to identifying the multiple strategies used by the SSIs and in the different contexts.

Jeanne Rose Century, a researcher at the Education Development Center, makes the case that as systemic reform calls for new roles for school administrators and teachers, the evaluator's role also is subject to change. As the trajectory of the field of evaluation during this century has evolved, so has the concept of evaluator from that of a technician to a more expanded role including advisor, collaborator, and coach. Century identified some specific roles for evaluators of systemic reform. Because of systemic reform's complexity, evaluators need to serve a multiplicity of roles and be versatile. The dynamic nature of systems forces those who are judging the value of reform to be flexible and to easily move in and out of specific roles. She argues that the goals for systemic reform fall within two domains: (1) improving educational practices and outcomes, and (2) building capacity. The success of systemic reform depends on instructional change and sustaining improvement through on-going reflection and reevaluation. An evaluator who identifies insufficient capacity within the system for it to achieve its goals may be in a position to provide technical assistance and may even be requested to do so by project leadership. Whereas in the past the tenet of independence and objectivity would inhibit an evaluator from providing technical assistance, the evaluator's understanding of the system may mean that he or she is in the best position to assist. The role an evaluator serves is shaped by many factors. In a systemic reform context, an evaluator's responsibility to a specific program or its staff may appropriately take precedence over the traditional constraints imposed on scientific objectivity.

Norma Dávila, University of Puerto and evaluator of the Puerto Rico Statewide Systemic Initiative, discussed alternative ways to measure student academic achievement within the new parameters of systemic

educational reforms. Achievement of challenging academic standards, as indicated by improved student academic achievement, is a central focus of the Puerto Rico SSI. The evaluation design of the Puerto Rico SSI was based on a participatory-research approach for evaluation, in general, and to assess student academic achievement, in particular. Evaluators triangulated their findings using multiple quantitative and qualitative methods. The evaluation employed three levels of assessment. Teachers trained in authentic assessment strategies used these assessment results to modify their practices and to monitor improvement in student academic achievement. The SSI staff developed a series of standards-based tests in science and mathematics closely aligned to classroom practices that were used to measure change in student achievement prior to teachers' participation in the SSI training and after completion of this training. An external measure was used as the third level of assessment. National Assessment of Educational Progress (NAEP) tests were adapted and translated to compare performance of students in schools participating in the SSI with those not participating.

Over time and as the SSI evolved, a more credible assessment was needed. The SSI staff, in an alliance with the College Entrance Examination Board (CEEB), developed assessments based on items from NAEP and the Third International Mathematics and Science Study (TIMSS) to measure achievement gains over one year. These tests were administered to all students from 377 schools in grades 4, 8, and 11. The scores from these assessments were scaled using the TIMSS scales so that the scores could be compared to international benchmarks. Similar, but not identical, assessment items were used to provide professional development to teachers on students' common misconceptions and how can they be corrected. Teachers in these sessions used the assessment items as a basis for examining their own performance. In addition to student assessment results, enrollment in higher education of students from SSI schools also

was being used as an outcome indicator. Although the SSI was comfortable with the original three levels of testing, the national exposure of the TIMSS reports since 1997 was a deciding factor in the decision to use an externally developed test. Using the publicly released items from NAEP and TIMSS was less expensive than developing their own tests, but required the expertise of CEEB. Dávila closed by observing the need for common metrics of student academic achievement that could be used across SSI sites.

Mark St. John, President of Inverness Research, drawing upon his training as a physicist, began by defining how he uses the terms *evaluation* and *systemic reform*. Evaluation refers to figuring out the value, the benefits, and the contributions that accrue from the public investment that is being made in the systemic initiatives. According to the theory of systemic reform, the instruction students receive—the quality of their learning experiences in schools—is directly shaped by the system—the political and institutional context—that surrounds classrooms. Any successful attempt to improve the quality of instruction must assume a systemic perspective in design and implementation. In education, as in other complex endeavors, there are many necessary yet not sufficient system supports that must be present if the system is to function well. Another important aspect of systemic reform is that the people who do the work of improving the system must be those living and working within the system. This means the people within the system must have the capacity and expertise to bring about intelligent change. Here capacity must consist of internal skills and knowledge, as well as access to external resources and expertise.

St. John identified “accountability misconceptions” that people have about how an education initiative should be judged that need to be confronted by evaluators. One is the “last input is the only input,” based on the conception that the quality of a teacher can be assessed by measuring the achievement of the teacher’s students. This ignores all of the educational and other experiences students

had prior to being in that teacher’s classroom. Another misconception is that improving only teacher preparation programs will improve the quality of teachers. This misconception also ignores the years of prior schooling the pre-service teachers have had. A third misconception is that the quality of a program or school can be judged by how high the test scores are. It is more accurate to identify good schools by those that add significant value to the knowledge and skills that students bring to schools. A fourth misconception is that a program, such as the SSI, is the only program in existence and can be studied in the absence of inputs from any other program. As such, clear effects can be attributed to each specific program. Evaluators need to be critics of unexamined and incorrect understandings.

St. John explained in more detail the difficulties that exist in establishing the value of NSF’s investment in systemic reform. One difficulty is the scale of the investment in relation to total education budget in the systems seeking change. Another is that there are many other factors that contribute to improved student learning. A third difficulty is that the impact of the investment on learning of any one student is very small by the time the investment is channeled through the many layers of the system from administration, curriculum, schools, teachers to classroom activities. A fourth difficulty is that the actual time required for NSF’s investment to have an impact might be longer than the funding period of five years.

Based on a study for the National Academy of Sciences, St. John developed a model depicting the relationship of key variables that could be used to judge the probability that a SSI would succeed. In this study, the single most important factor was the quality, expertise, commitment, and political power of the leadership. Other important factors were the knowledge and expertise that exist within the reform itself (design), policy and reform infrastructure, discretionary funding that can be allocated specifically towards reform, and the political and public demand for reform. For systemic reform to be effective, these factors have to overcome barriers that include scale of the

system, political "cross currents," severe financial problems, instability and turbulence in the system, other reforms, and competing priorities. Based on this model, what NSF should be held accountable for should be the degree to which its investments build the state or district capacity for initiating, and sustaining, reform. NSF should not be held

accountable for what a state or district does with this capacity. He advises that it is better to document the contributions the systemic initiatives are making to increase the capacity of teachers and others than to argue that they are directly causing increased student achievement in the short term.

CRITICAL ELEMENTS OF AN EVALUATION OF SYSTEMIC REFORM

Patrick M. Shields, Andrew A. Zucker, and Nancy E. Adelman
SRI International

In 1990, the National Science Foundation (NSF) launched the Statewide Systemic Initiative (SSI) program to help states undertake comprehensive and coordinated reforms of mathematics, science, and technology education. Between 1991 and 1993, NSF signed five-year cooperative agreements with 25 states and the Commonwealth of Puerto Rico to carry out standards-based systemic reform throughout their jurisdictions.

To assess the extent to which states have undertaken the kinds of changes envisioned by NSF, and to examine the efficacy of different SSI strategies, the Foundation contracted with SRI International to conduct a national evaluation of the program. This paper reviews the framework the evaluation team used for assessing the progress of the SSIs, outlines the evaluation methodology, and reflects on a number of challenges involved in evaluating systemic reform.⁴

A Framework for Assessing State Strategies

The concept of systemic reform has been outlined by Smith and O'Day (1991) and elaborated numerous times since then (see Clune, 1993; Fuhrman & Massell, 1992; Fuhrman, 1993; Vinovskis, 1996). The essence of the concept is that ambitious standards for student learning should form the basis for the alignment of all policies, practices, and resources throughout the educational system. Fundamental to the concept is that ambitious goals apply to all students, not just those destined for professional careers (O'Day & Smith, 1993).

To guide the evaluation of the SSI program, the evaluation team developed a

conceptual model of systemic reform, shown in Exhibit 1, specifying the key components of the educational system that needed to be reformed in concert. The Exhibit shows SSI activities or investments moving in two related but distinct channels. One set of investments has been made for activities relatively close to students and teachers, including support by the SSIs for professional development. A second set of activities has focused on activities more distant from classrooms, such as the development and dissemination of state curriculum frameworks. Because systemic reform aims to change both student outcomes and the education system itself, both sets of activities have been important. However, different SSIs have supported widely varying combinations of strategies to effect changes at different levels of the education system (see Zucker & Shields, 1997). Other key features of the model are as follows.

The Top of the Model: Students, Teachers, Classrooms, and Schools

By placing student outcomes at the apex of the figure, the model emphasizes that the overarching goal of systemic reform is to raise student achievement, increase students' interest and enrollment in challenging courses, and otherwise *improve education outcomes for young people*. Improvements in student learning rest on *improved classroom experiences*. Such experiences are characterized by active student engagement with real-world scientific and mathematical problems, critical inquiry into a limited set of topics, and opportunities for actual scientific thinking and discourse (CSMEE, 1997; Project 2061, 1993). In contrast to the typical American school, classrooms that provide such experiences are marked by less teacher-directed instruction, more student-student interaction, the flexible organization of space

⁴ This paper is based on a series of reports produced for the evaluation, references to which can be found at the end of this paper.

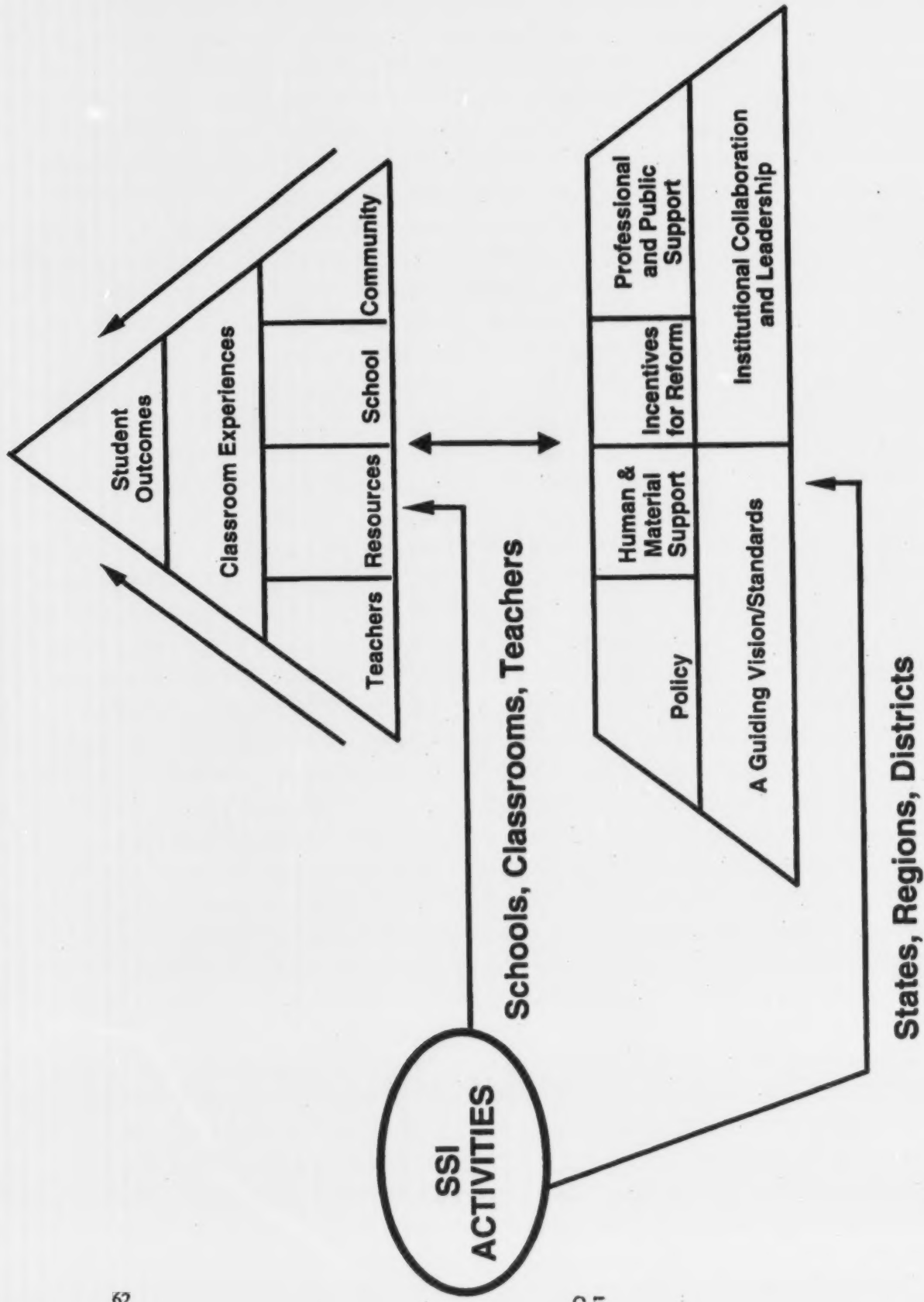


Exhibit 1: A MODEL OF SYSTEMIC REFORM

and time in line with the specific learning goals at hand, and regular constructive feedback to students based on their performance on actual mathematics and science tasks (That-p & Gallimore, 1989).

The creation of such classrooms, the model continues, calls for teachers *with a new set of skills, resources, and knowledge*. Teachers must have a thorough command of their subject matter—an especially challenging task in mathematics and science, particularly at the elementary level (Cohen & Hill, 1997). Teachers must understand how students learn and how to structure learning opportunities to capitalize on students' knowledge and learning styles (Darling-Hammond, 1996; National Commission on Teaching and America's Future, 1996). Perhaps most importantly, they must believe that all their students can master challenging content.

Beyond content knowledge and pedagogical skills, teachers and their students must have access to appropriate tools and instructional materials. They need classroom technology (e.g., lab equipment, graphing calculators) and high-quality instructional materials. Access to appropriate curricula is particularly important because the challenge of creating inquiry-centered classrooms is already so daunting that without good curricula, teachers are faced with the prospect of creating their own materials while simultaneously struggling to change their own practice (Adelman, 1997; Zucker, 1997).

The provision of needed material resources, as well as the time teachers need to plan and assess the teaching and learning in their classrooms, calls for associated 'changes in the culture and organization of schooling. Fullan (1996) uses the term "re-culturing" to refer to fundamental shifts in a school away from traditional norms structured by bureaucratic roles to a philosophy where student attainment of high standards is the central concern of all staff. Restructuring refers to the reorganization of standard operating procedures, especially time and the use of space, to promote

student and teacher learning. From this perspective, schools that are supportive of teachers creating effective classrooms are characterized as learning organizations. Teachers have time away from children to interact and reflect with their peers; resources are allocated to optimize learning; and the scheduling of class periods as well as the grouping of students is flexible and driven by learning goals (Elmore & Associates, 1990). Such schools also reach out to others because they require the support and buy-in of parents and the local community. Parent and community support is especially important when fundamental shifts in classroom practice are envisioned, as promoted by systemic reform (Shields & Knapp, 1997).

The Base of the Model: Districts, Regions, and States

To support reforms at the school and classroom levels on any scale requires coordinated and coherent reforms at the levels of states, regions, and districts. Of paramount importance is the alignment of policies at the state and local levels. The misalignment of traditional **basic-skills**-oriented, norm-referenced tests with new and ambitious goals for student learning was one of the fundamental concerns of the proponents of systemic reform (Smith & O'Day, 1991). More coherent and robust policies are needed to send consistent messages to educators and the public about what is valued. Beyond assessments and frameworks, there are a host of policies under the control of either the state or local districts, depending on political traditions, that influence who ends up in classrooms, how teachers teach, and what support teachers receive.

Beyond policy, there is the need for building an infrastructure at the district, region, and state levels that will provide human and material support required for school and classroom reforms. The task faced by district and state administrators is no less challenging than that confronting

classroom teachers, and the "system" that holds together district and state efforts is just as disjointed as the typical school. Systemic reform calls for districts and states to jettison their traditional role as regulators of local practice and assume the new role of technical assistants to schools. They have to understand, and be willing to address, the resource allocation, professional development, and organizational issues raised by the reforms (see Spillane & Tompson, 1997).

A third factor that districts and states need to address is incentives for reform. Changing practice requires extra time and effort by teachers-time for learning, time for redesign-and it entails some risk, including the possibility of inadequate performance; negative reactions from colleagues, students, or parents; or lower achievement. So teachers must be highly motivated to undertake changes; they must have compelling reasons for taking on the work and the possible risks. Persuading large numbers of teachers and school administrators to engage in the work of reform requires the alignment of existing incentives with reforms, the elimination of disincentives, and sometimes the creation of new or additional incentives. Guidance mechanisms such as state standards, state and local assessments, and personnel evaluation criteria are all critical parts of the incentive structure affecting classroom practice. Many reformers also call for strong accountability systems that include public release of student outcomes and clear rewards and sanctions (David, 1990).

The fourth reform task at the state and district levels involves building professional and public support for the reform agenda. Systemic reform requires widespread public acceptance and support. The public may sometimes appear apathetic about instructional reforms, but changes in the classroom that depart from the public's conceptions of "real" school will quickly galvanize parents if their support has not been obtained in advance. In democratically controlled school systems with weak

professional structures, classroom practice is not determined solely by professionals. Instead, teaching practice is subject to close public scrutiny by parents and community members, and changes in practice require public acceptance, as well as formal approval by local boards.

A well-specified vision of student learning goals forms the basic premise of all versions of systemic reform. The argument is simple: coherence and alignment in the educational system must be guided by a shared understanding of what we want students to learn. In the early writing on systemic reform, this vision was likely to be specified in curriculum frameworks-again based on the experience of California in the 1980s (Smith & O'Day, 1991). Throughout the mid-1990s, as states tried out many of the ideas of systemic reform, curricular frameworks were replaced by state standards as the key vehicle for communicating a vision of high-quality instruction and learning. In fact, the term *systemic reform* was often replaced in the literature by the term "standards-based reform" (David, Shields, Young, Glenn, & Humphrey, 1997).

High-level leadership and collaboration among key institutions at both the state and local levels are required to help assure the legitimacy of the reform vision and thus its political power to guide shifts in policy and practice, as well as to motivate the concentration of resources needed for reform. The task of fundamental reform is both technical and political. Technically, it requires collaboration among the best minds-to set standards, realign assessment systems, restructure incentive systems, and build an appropriate infrastructure to support the reform effort. Politically, it requires the will to agree on a single set of learning outcomes, to establish appropriate accountability mechanisms, to build public support, and to garner the necessary fiscal resources. Achievement of both the technical and political tasks of reform is impossible without the buy-in and support of the top leadership.

Systemic Reform: A Summary

In summary, the model of systemic reform we have outlined follows a deceptively simple logic. Based on clear standards for what students should know and be able to do, and with the support of the key leadership, states and districts must align policy, build the capacity to provide schools and teachers with needed human and material support, restructure incentive systems, and build professional and public support for the reform agenda. These actions in turn are meant to provide the support needed to help increase teachers' capacity to implement the reform vision with access to appropriate material, within schools organized to support their efforts, and with the support of parents and community. In such contexts, the argument continues, reformed classroom practice can occur and student learning will increase.

The Evaluation Methodology

The evaluation is based on data collected from a wide variety of sources. Three sources were most important. First, quantitative data were gathered annually from the principal investigators in each SSI. In addition, the evaluation team conducted repeated site visits in every SSI. Finally, secondary data analysis included careful study and, in some cases, reanalysis of data sets gathered by many of the SSIs as part of their ongoing efforts to assess progress toward reaching their goals.

The evaluation included a set of 12 detailed case studies, for SSIs in Arkansas, California, Connecticut, Delaware, Kentucky, Louisiana, Maine, Michigan, Montana, New York, Vermont, and Virginia. The time on-site in each case study state averaged about 50 person-days. Site visiting took place both during the school year and in the summer. More than two dozen districts in the case study states were described in detail by the evaluation team (but the written descriptions were not published), as well as more than three dozen schools.

In the thirteen non-case-study states, the time on-site averaged about six person-days per SSI, and, again, a very large amount of information was gathered and analyzed about each of them. By design, these visits were briefer, were less frequent, and typically involved only a single evaluator. Written descriptions were not published; however, they averaged about 25 pages single-spaced for each of the non-case-study SSIs.

In all the states, on-site visits were supplemented with telephone interviews, in-person interviews at periodic meetings of the SSI principal investigators and project directors, and extensive document analyses. Documents reviewed included monitoring reports about each SSI that were produced by Abt Associates, multiple documents written by each SSI (such as annual reports to NSF), and reports of a number of evaluations conducted for specific SSIs. The latter were especially useful for developing two of the evaluation reports that focus on what selected SSIs learned about the impacts of their activities on teachers' classroom practices and on student achievement. As necessary, information about particular states was also updated via telephone or e-mail to be sure information in each report was current.

Reflections on the SSI Evaluation

The evaluation of statewide systemic initiatives in 26 states presented a massive challenge: we were essentially attempting to track the progress of 26 distinct efforts to reform the entire system of education and to then make overall judgements of the success of those efforts in the aggregate. In undertaking this daunting task, we learned some lessons about what we did right and about where future evaluations can be strengthened.

What We Did Right

Our overall approach to the evaluation, building on a clear model of systemic reform, studying entire state systems, and

identifying specific reform strategies proved useful in meeting the challenges presented in this evaluation.

A Model of Systemic Reform as an Evaluation Framework. We began the evaluation of the SSIs with a cruder version of the model presented earlier in this paper. This model served us well in identifying a set of components that should be included in system-wide reform and it helped us to develop hypotheses about the interrelationship of these components that could be tested against the empirical data from the sites. In general, we found that the model provided the appropriate categories and relationships among those to describe, analyze and assess the activities of the individual SSIs. The model also served to facilitate cross-case comparisons and to underscore areas where the SSIs, taken as a whole, had more or less impact on the entire system of education.

Conducting In-Depth Case Studies of Entire State Systems. Systemic reform by definition is meant to involve an entire state system. Each of the participating systems presented a unique set of **circumstances**—not only in terms of demographics, geography, political culture and fiscal resources, but also in terms of ongoing reform efforts into which the SSI fit. Understanding the progress of the SSI required understanding the evolution of educational reform in the state as a whole. In short, the SSI could not be studied as a “project,” separate from other reform initiatives. Consequently, we chose to conduct in-depth case studies in a range of states in order to tell the full reform story in those contexts. These provide detailed descriptions and analyses of the progress of individual SSIs within the context of mathematics and science reform in their states.

Identification of Specific State Strategies. Each of the SSIs’ total reform efforts consisted of a set of related change strategies. We identified eight of these:

- Supporting teacher professional development

- Developing, disseminating, or adopting instructional materials
- Supporting model schools
- Aligning state policy
- Creating an infrastructure for capacity building
- Funding local systemic initiatives
- Reforming higher education and the preparation of teachers
- Mobilizing public and professional opinion.

Although we were not able to identify a small number of “models” or “types” of systemic reform with which to categorize and assess the SSIs, the identification of these eight strategies facilitated cross-site analysis and allowed us to assess the strengths and weaknesses of focusing on specific components of the system.

Taken together, the use of a comprehensive framework, the focus on whole state systems, and the identification of specific SSI strategies allowed us to provide accurate pictures of individual state’s progress while making cross-site conclusions about the relative efficacy of different SSIs.

The Jury Is Still Deliberating

The evaluation did not meet every goal we set out for ourselves. In retrospect some of these goals may not have been realistic or even possible. Yet, as researchers and evaluators consider future work, it is worthwhile to reflect on some of these issues.

Reliance on the SSIs for Statewide Impact Data. The goal of systemic reform is to improve teaching and learning. We found it infeasible to collect comparable classroom and student impact data across all 26 states, the thousands of schools, and tens of thousands of teachers involved in the reforms. Early on in the evaluation, we developed and piloted a teacher survey that sought comparable classroom data. But we found it impossible to calibrate an instrument that was sensitive enough to

gauge the kinds of teaching practice we were interested in and that could be used across multiple SSIs. The development and implementation of multiple surveys for many different SSIs was deemed too expensive. As a result, we relied on the data from the case studies, which always involved a small subset of classrooms, and on data from the SSIs' internal evaluations. Because of the unevenness of the internal evaluations, we were left with very uneven data on the classroom impact.

Much the same can be said regarding student learning. NSF made a decision early on in the evaluation not to support a common assessment instrument across sites nor to require the use of a specific instrument. During the course of the SSIs, most states changed testing policies at least once and many never implemented a test designed to assess the type of learning the SSIs sought to promote. As a result, we were left with no data on student achievement from a number of SSIs and non-comparable data where they existed at all.

We did end up producing reports on both student achievement and classroom impacts, but each was based on data from selected states and neither provided quantitative cross-site analyses using comparable measures of progress.

Creating a Report Card. NSF invested heavily in the SSIs and in the evaluation of their progress. At different times during the evaluation, the Foundation sought a concise summary of the relative progress of the states. We found this task quite challenging as it required us to reduce the complex stories of 26 initiatives operating in very different reform contexts to simple scores or rankings. In a compromise with NSF, we ultimately produced an internal memo to the Foundation in which we scored the progress of the 26 SSIs for each of the eight strategies described earlier in this paper as well as for a set of crosscutting dimensions. The ultimate analysis resulting from this exercise will be published in a forthcoming paper (Adelman, Shields, & Zucker, forthcoming)-although the state-by-state ranking will not be made public.

Whether a reliable "report card" could have been produced remains an open question. Efforts by the American Federation of Teachers, *Education Week*, and others to assess components of state reform efforts have proven quite unreliable.

Identifying and Assessing "Models" of Systemic Reform. From the beginning of the evaluation, NSF encouraged the evaluation team to identify a small set of models of systemic reform. The goal was to identify a limited set of approaches to systemic reform and then to assess their relative efficacy. The argument was that while there was certainly more than one way to reform a system of mathematics and science education, there were certainly less than 26 ways to do so.

In the end, we made a great deal of progress toward the goal of identifying models, but never quite reached it. As discussed earlier, we did identify a finite set of strategies for achieving systemic change. We identified which SSIs employed these strategies, we described which SSIs relied heavily or even primarily on one or another strategy, and assessed the degree to which individual SSIs succeeded in implementing an individual strategy. We were also able to assess the degree to which the SSIs used various strategies in more or less comprehensive approaches to full system reform. Yet, because each of the SSIs employed multiple strategies in different combinations and within very different contexts, we were not able to identify a small set of model approaches to systemic reform.

References*

- *Adelman, N. E., Shields, P. M., Goertz, M. E., Zucker, A. A., & Corcoran, T. B. (1998). *Doing systemic reform: The experiences of the SSI states*. Menlo Park, CA: SRI International.

* Publications produced as part of the evaluation of NSF's Statewide Systemic Initiatives program are marked with an asterisk.

- Adelman, N. E., & Walking Eagle, K. P. (1997). Teachers, time, and school reform. In A. Hargreaves (Ed.), *Rethinking educational change with heart and mind: 1997 ASCD Yearbook*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Blank, R. K., Langesen, D., Bush, M., Sardina, S., Pechman, E., & Goldstein, D. (1997). *Mathematics and science content standards and curriculum frameworks: States' progress on development and implementation*. Washington, DC: Council of Chief State School Officers.
- Blank, R. K., Hemphill, C., Sardina, S. L., Langesen, D., & Braithwaite, B. (1995). *State education policies on K-12 curriculum, student assessment, and teacher certification: 1995*. Washington, DC: Council of Chief State School Officers.
- *Breckenridge, J. S., Goldstein, D. S., & Zucker, A. A. (1996). *The impact on students of the SSI Program: A pilot study of the impacts of the Louisiana and Montana SSIs (draft)*. Menlo Park, CA: SRI International.
- Center for Science, Mathematics, and Engineering Education, National Research Council. (1998). *Every child a scientist: Achieving scientific literacy for all*. Washington, DC: National Academy Press.
- Clune, W. H. (1993). The best path to systemic educational policy: Standard/centralized or differentiated/decentralized? *Educational Evaluation and Policy Analysis*, 3 (15), 233-254.
- Cohen, D. K., & Hill, H. C. (1998, January). *CPRE policy briefs. State policy and classroom performance: Mathematics reform in California*. Philadelphia: Consortium for Policy Research in Education.
- *Consortium for Policy Research in Education. (1995, May). *CPRE policy briefs. Reforming science, mathematics, and technology education: NSF's State Systemic Initiatives*. New Brunswick, NJ: Rutgers University. Author.
- *Consortium for Policy Research in Education. (1995, July). *CPRE policy briefs. Tracking student achievement in science and math: The promise of state assessment systems*. New Brunswick, NJ: Rutgers. Author.
- *Corcoran, T. B., Shields, P. M., & Zucker, A. A. (1998). *Evaluation of NSF's Statewide Systemic Initiatives (SSZ) Program: The SSIs and professional development for teachers*. Menlo Park, CA: SRI International.
- Darling-Hammond, L. H. (1996). Teaching and knowledge. In J. Sikula (Ed.), *Handbook of research on teacher education* (2nd ed.). New York: Association of Teacher Educators.
- David, J. (1990). *Results in education: State actions to restructure schools: First steps*. Washington, DC: National Governors' Association.
- David, J. L., Shields, P. M., Young, V. M., Glenn, B. C., & Humphrey, D. C. (1997, October). *Pew network for standards-based systemic reform: Year one evaluation report*. Menlo Park, CA: SRI International.
- Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, 66 (1), 1-26.
- Elmore, R. F., & Associates. (1990). *Restructuring schools: The next generation of educational reform*. San Francisco: Jossey-Bass.
- Fuhrman, S. H., & Massell, D. (1992). *Issues and strategies in systemic reform*. New Brunswick, NJ: Rutgers University, CPRE.
- Fuhrman, S. H. (1997). *What has been learned about systemic reform?* Madison, WI: University of Wisconsin, National Institute for Science Education.
- Fullan, M. G. (1996, February). Turning systemic thinking on its head. *Phi Delta Kappan*, 77 (6), 420-423.
- Humphrey, D.C., & Shields, P.M. (1996). *A review of mathematics and science curriculum frameworks*. Menlo Park, CA: SRI International.

- Humphrey, D. C., Anderson, L., Marsh, J., Marder, C., & Shields, P. M. (1997). *Eisenhower Mathematics and Science State Curriculum Frameworks Projects: Final evaluation report*. Menlo Park, CA: SRI International.
- Knapp, M. S., & Associates. (1995). *Teaching for meaning in high-poverty classrooms*. New York: Teachers College Press.
- Knapp, M. S. (1996). *Between systemic reforms and the mathematics and science classroom*. Madison, WI: University of Wisconsin, National Institute for Science Education.
- *Laguarda, K. G., Breckenridge, J. S., & Hightower, A. M. (1994). *Assessment programs in the Statewide Systemic Initiatives (SSI) states: Using student achievement data to evaluate the SSZ*. Washington, DC: Policy Studies Associates.
- *Laguarda, K. G. (1998). *Assessing the SSIs' Impacts on Student Achievement: An Imperfect Science*. Menlo Park, CA: SRI International.
- Little, J. W. et al. (1987). *Staff Development in California*. San Francisco, CA: Far West Laboratory.
- Massell, D., Kirst, M., & Hoppe, M. (1997). *CPRE Policy Briefs. Persistence and Change: Standards-Based Systemic Reform in Nine States*. Philadelphia: University of Pennsylvania (Consortium for Policy Research in Education).
- National Commission on Teaching and America's Future. (1996). *What Matters Most: Teaching for America's Future*. New York: Author.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (1993). *Professional Standards for Teaching Mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: Author.
- National Research Council. (1995). *National science education standards*. Washington, DC: National Academy Press.
- National Science Foundation. (No date). *The National Science Foundation's Systemic Initiatives*. Arlington, VA: Author.
- O'Day, J. A., & Smith, M. S. (1993). Systemic reform educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 250-312). San Francisco: Jossey-Bass.
- Project 2061, American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- Purkey, S., & Smith, M. S. (1983). Effective schools: A review. *Elementary School Journal*, 83 (4), 427-452.
- Resnick, L. B. (1987). *Education and learning to think*. Washington, DC: National Academy Press.
- Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of innovations: A cross-cultural approach*. New York: Free Press.
- *Shields, P. M., Corcoran, T. B., & Zucker, A. A. (1994). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: First Year Report. Volume I: Technical Report*. Washington, DC: National Science Foundation.
- Shields, P. M., & Knapp, M. S. (1997, December). The promise and limits of school-based reform. *Phi Delta Kappan*, 288-294.
- *Shields, P. M., Marsh, J. A., & Adelman, N. E. (1998). *Evaluation of NSF's Statewide Systemic Initiatives (SSI) Program: The SSIs' Impacts on Classroom Practice*. Menlo Park, CA: SRI International.
- *Shields, P. M., Marsh, J. A., Marder, C., & Wilson, C. L. (1998). A case study of California's SSI (CAMS), 1992-1997. In P. M. Shields & A. A. Zucker (Eds.), *SSI Case Studies, Cohort 2: California, Kentucky, Maine, Michigan, Vermont,*

- and Virginia. Menlo Park, CA: SRI International.
- *Shields, P. M., & Zucker, A. A. (Eds.). (1998). *SSI Case Studies, Cohort 2: California, Kentucky, Maine, Michigan, Vermont, and Virginia*. Menlo Park, CA: SRI International.
- Smith, M., & O'Day, J. (1991). Systemic school reform. In S. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing*. Bristol, PA: Falmer Press.
- Spillane, J. P., & Thompson, C. L. (1997). Reconstructing conceptions of local capacity: The local education agency's capacity for ambitious instructional reform. *Educational Evaluation and Policy Analysis*, 19 (2), 185-203.
- Tharp, R., & Gallimore, R. (1989). *Rousing minds to life*. Oxford: Oxford University Press.
- Vinovskis, M. A. (1996). An analysis of the concept and uses of systemic educational reform. *American Educational Research Journal*, 33 (1), 53-85.
- Zucker, A. A. (1997). *Reflections on state efforts to improve mathematics and science education in light of findings from TIMSS*. Menlo Park, CA: SRI International.
- *Zucker, A. A., Shields, P. M., Adelman, N., & Powell, J. (1995). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSZ) Program: Second Year Report. Part I: Cross-Cutting Themes*. Washington, DC: National Science Foundation.
- *Zucker, A. A., & Shields, P. M. (Eds.). (1995). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: Second-Year Case Studies: Connecticut, Delaware, and Montana*. Menlo Park, CA: SRI International.
- *Zucker, A. A., & Shields, P. M. (1997). *SSI strategies for reform: Preliminary findings from the evaluation of NSF's SSI Program*. Menlo Park, CA: SRI International.
- *Zucker, A. A., & Shields, P. M. (Eds.). (1998). *SSI Case Studies, Cohort 1: Connecticut, Delaware, Louisiana, and Montana*. Menlo Park, CA: SRI International.
- *Zucker, A. A., & Shields, P. M. (Eds.). (1998). *SSI Case Studies, Cohort 3: Arkansas and New York*. Menlo Park, CA: SRI International.

EVALUATORS' ROLES: WALKING THE LINE BETWEEN JUDGE AND CONSULTANT

Jeanne Rose Century
Education Development Center

The practice of educational evaluation has recent historical roots in the early part of the century when intelligence tests and the notion of "scientific management" of education were first developed. The principles that grew out of this movement, such as using carefully crafted tests to find "scientific" solutions to educational problems, exerted influence on what today has become the educational evaluation enterprise. Their remnants are evident in many of today's evaluations in which evaluators serve as "judges" and gather quantitative data on student, teacher and school performance in order to draw conclusions about program effectiveness and worth.

This role of "judge" is a necessary, frequent part of many evaluation plans. But as new theories about evaluation practice have evolved over the last few decades so have new ideas about evaluators' roles emerged to expand and complement this basic function. Education reformers, researchers, evaluators and funders have debated these roles and the "place" of the evaluator in a reform. They have asked whether evaluators should be internal or external to programs; whether reporting and feedback should be **summative** or formative; and whether to use qualitative or quantitative methodologies. Madaus and Kellaghan capture this debate by stating that

the emphasis that evaluation and assessment have received and the form they have taken at different points in history reflect differences in the nature of education and the determinants of school achievement, the importance of accountability, and the purpose of evaluation. (Madaus & Kellaghan, 1992, p. 121)

Now, the systemic reform movement again stimulates the development of new educational theory. In turn, evaluators must consider that this is also a time of change for the evaluation enterprise. Systemic reform calls for new roles for project leaders, school administrators and teachers; it seems the role in the evaluator is likely to change as well.

Evolution of Evaluation Roles

There is little, if any, consensus as to what evaluators' roles should be, whose values should be represented in evaluation, and what questions evaluators should ask (Shadish et al., 1991). While a program's goals, purposes and context ultimately determine the answers to these questions, over the last forty years, theories of evaluation have developed which influence this debate. Evaluation has become a more prominent enterprise in the education endeavor, bringing with it new descriptions of evaluators' roles and functions.

In *Foundations of Program Evaluation: Theories of Practice*, for example, the authors describe the evolution of new ideas about evaluation in three stages (Shadish et al., 1991). The first stage was rooted in a scientific approach to finding successful solutions to social problems. The second stage grew in the 1970s and represented an interest in departing from traditional practices to create approaches to evaluation that were more practical and would be of more use to the programs. The third stage of evaluation theory was focused on integrating all of the methodologies and strategies that had come before into a more "coherent" approach to evaluation (Shadish et al., 1991).

Guba and Lincoln (1989) also developed a categorization scheme for evaluation. They describe four "generations." The first has

evaluators in the role of "technician," in which they are familiar with an existing set of measurements and identify the most appropriate for the task at hand. The second generation places evaluators in the role of "describer," in which they extend the measurement role to include "chronic[ing] of program strengths and weaknesses." The third generation casts evaluators as "judges," who assess the worth of a program, and the "fourth generation" evaluator is one who retains each of the previous roles, but adds new ones such as: collaborator, learner/teacher, reality shaper, mediator and change agent (Guba & Lincoln, 1989; O'Sullivan, 1995).

Shadish's third stage and Guba and Lincoln's fourth "generation" place the theoretical discussion about evaluation on a trajectory that seems compatible with the evolution of systemic reform. Just as theory of educational change is evolving to accommodate systemic approaches, so are discussions about evaluators' roles evolving to encompass a wider range of responsibilities and purposes. In the early part of this decade, for example, Beswick wrote that the role of the evaluator was moving from what one might describe as technical roles to more political and advisory roles (Beswick, 1990). Similarly, others suggested that education reform efforts needed evaluators to function as coaches or collaborators in order to most effectively respond to the demands for accountability and impact (McColskey, 1995). Now, as systemic reform becomes more widespread, opportunities for, and interest in such non-conventional roles grows with it.

Evaluators' Roles in Systemic Reform

Just as there is no single approach to implementation of systemic reform, there is no single model for evaluating it. There do, however, seem to be some common themes regarding evaluation roles in systemic reform that are likely to influence how evaluators develop evaluation plans and strategies. First, evaluators can expect to play multiple roles. Building from Shadish's third stage of "coherence," or Guba and Lincoln's fourth generation, evaluators of systemic reform

must have a versatility that will allow them to serve a variety of needs. Systemic reform is quite complex and involves multiple stakeholders. Consequently, evaluators may need to shift roles to best match the various targeted areas of study within the systemic reform and to best accommodate the interests and needs of the client at any particular time.

Second, hand in hand with the complexity of systemic reform is the dynamic, fluctuating nature of the systemic endeavor. This suggests that evaluators have to do more than accommodate different roles at different times, but that they also need to move in and out of those roles in a flexible manner. Depending on how the evaluation is organized, individual evaluators may work within a clearly defined set of roles, or they may need to play multiple roles somewhat simultaneously. Every set of evaluation roles for a systemic reform effort will be different; each evaluation effort will have a somewhat different purpose and goal. There is no predicting which roles will always have to be played when, but it is important that those participating in the evaluation together have a palette of roles which can allow them to best meet the needs of the reform and fulfill the goals and purposes of the evaluation.

Nearly 15 years ago, Maurice Eash recognized the potential for change in evaluation. He wrote:

... our relationship with the client has changed drastically. . . . The process has moved from one that was set very much in advance to one that requires a continuing interface with the client and is largely evolutionary. . . . As for the future, I believe the following well-established trends will continue: a) close interaction of the evaluator and client throughout the life of a project using multilevel evaluation, b) evolving rather than fixed evaluation designs and c) addressing of design questions and findings to numerous interest groups by giving equal attention to the contextual politics involved as well as the technical demands." (Eash, 1985, p. 252)

This description seems to capture some of the issues underlying evaluation of systemic reform quite well. The evaluator-client relationship may no longer be one that is strictly formal and confined to conventional roles. Rather, it is likely to adapt to the needs of the reform, the evolving purposes and goals for the evaluation and the specific client/audience at any particular time.

The range of roles, then, that evaluators might be called upon to play is expanding. The list is long and includes: collaborator, documentor, critical friend, advocate, teammate, coach and change agent to name a few. Discussion of some of these roles has been continued over the years, but has been largely ignored until the emergence of systemic reform. Other roles are new with the arrival of the systemic scheme, and still others have attained greater significance and meaning in the systemic arena.

In order to understand how these and other roles are necessary to address the needs and purposes of a systemic program, one must look closely at the organizing goals of systemic reform efforts. One can argue that these goals fall within two domains. One domain, much like conventional reforms, focuses on improving educational practices and outcomes. The second domain focuses on capacity building and is tied to the nature of systemic reform as a continuous endeavor (Century, 1997). Therefore, in addition to the structural and instructional changes that a mathematics or science reform puts in place (the first domain) a systemic reform must also focus on establishing policies, practices, and a culture and environment that will support continuous positive development in the future (the second domain). Success of systemic reform then, includes both, establishment and maintenance of new instructional changes as well as an enduring ability to reflect on, reevaluate, and improve both new and old practices once they are in place.

There are three categories of roles that evaluators can play when responding to these two domains: evaluation roles-those roles that evaluators have played in the past but that take on increased importance and significance in the context of systemic reform; systemic

perspective roles-those roles that are new or uniquely significant in the context of systemic reform; and technical assistance roles-roles that are typically played by technical assistants but can be appropriate for evaluators in the context of systemic reform (Century, 1997). Placement of a role in one specific category does exclude it from others; some roles may overlap into two or even all three categories.

The two domains of systemic reform goals link directly to the traditional evaluation roles and the technical assistance or consultant roles respectively ("systemic perspective" roles won't be addressed here). For example, an evaluator addresses the first domain (improvement in educational practices and outcomes) through judging gains in the education change process. Evaluators might ask questions such as: "Are new instructional materials in place?" "Has professional development improved?" "Do student assessments reflect the changes in the curriculum?" and "Is there improvement in student performance?" Evaluators identify the presence of changes in these various aspects of the educational program and the extent to which those changes are of high quality. Then they can make judgments about the success of the reform.

This might be where evaluations of more conventional programs stop. In systemic reform, however, evaluators also can turn to the second domain of goals: capacity building. In doing so, they need to consider whether there is sufficient capacity in the system to sustain continuous growth. This still aligns with the role of judge, but brings evaluators to the edge of what some would consider unacceptable practice. If evaluators see that there is insufficient capacity in the system, they are confronted with a challenge: whether or not to cross the line from judging capacity and take on a consultant role to assist in building the capacity. Some evaluators have found that this question is answered for them. Reluctantly, or even perhaps unwillingly or unknowingly they find themselves responding to the needs of the client by crossing what is sometimes a blurry line between evaluator and consultant.

Should an Evaluator Assume the Role of the Consultant or Technical Assistant?

In writing about providing assistance to third world nations, Harari describes the technical assistance expert as one who is "an instrument of communication between two worlds . . . his official vocation is to foster the development of the country to which he is sent by virtue of the work he does" (Harari, 1974). Some evaluators working in the field in systemic reform would concur that even though they don't set out to play technical assistance roles that fit this description, project leaders and participants implicitly ask them to do so. For many, engaging in technical assistance roles is inappropriate. Such actions compromise what is, in their eyes, the necessary objectivity of the evaluator and the evaluator's ability to be a fair judge. These concerns weigh heavily in a field that has worked to develop a careful set of methodologies and strives for the credibility that results from adherence to these methods and the standards that accompany them.

And yet, as mentioned above, systemic reform seems to call for a redefining of roles of all participants. Like state and district administrators, teachers, and other stakeholders, evaluators and technical assistants "have been taught and have come to believe a set of conventions about what their titles mean and what actions are within and outside of their bounds" (Century, 1997). When exploring the new realm of evaluation of systemic reform then, they may find it necessary to question some of the standing assumptions about roles and consider the appropriateness of those roles in a different light.

The suggestion that evaluators begin to act as technical assistants is not new. Writing about utilization of evaluation findings, for example, Eash noted that the most significant factor contributing to increased use of evaluation findings was when "evaluation assumed an 'educative' approach to the client throughout the process" (Eash, 1985). More recently, writing specifically about mathematics and science curriculum

development, O'Sullivan noted that in addition to more conventional evaluation roles, "funding agencies and other program sponsors are requesting that evaluators provide technical assistance to programs that may require on-going, active evaluation assistance through the project development process" (O'Sullivan, 1995).

Ultimately, an evaluator's choice of role is influenced by many different factors including the general purposes and goals of the evaluation, the relationship with the client, the identified audience, as well as the personal and professional considerations of the evaluator, including experience, style, and training (Alkin et al., 1979). Some roles, however, evolve without intention, influenced by the emerging shape of the systemic reform and its evaluation. Evaluators may find that as the reform progresses, needs emerge which can be met only through assuming some unanticipated roles. In conventional reforms, evaluation responsibilities stop at directing attention to the needs of the project, not responding to them. In systemic reform settings, (or even in what Shadish refers to as policy or program settings) the evaluator may feel that "responsibility to a specific program, its staff and stakeholders often takes precedence over traditional role behaviors of scientists" (Shadish et al., 1991) and he/she may feel compelled to take action in a new, unanticipated role.

Clearly, changes in evaluation that include consultant roles place evaluators beyond what many evaluators would consider acceptable boundaries between the evaluator and the evaluand. They call for increased involvement that goes beyond the limits outlined by some, while staying well within the limits set by others. There is a line to be drawn somewhere between the evaluation endeavor and the systemic reform; that line may fall in a different place, for different reform efforts and for different evaluators. While the boundary line will never be completely clear, evaluators need to clarify for themselves their best understandings of where that line is drawn for each role and circumstance.

Redefining Conventions of Evaluation

Underlying much of this discussion about role are deeper questions about the implications these roles have for current understandings of the credibility, validity, and objectivity of the evaluator. Objectivity and credibility are typically linked in that without objectivity, an evaluator can not be credible. Similarly, the methodologies and relationships suggested by some of the roles described here threaten current understandings of the validity of the evaluation. The specific implications for each of these touchstones of evaluation are far too complex to address here. However, it is important that evaluators recognize that this may be a time when the field generates new meanings for these words, at least as they exist in the context of evaluation of systemic reform efforts.

This paper has been adapted from chapters written for the NISE book, *Evaluation of Systemic Reform in Mathematics and Science*, which is under preparation.

References

- Alkin, M. (1972). Evaluation theory development. In C. Weiss (Ed.), *Evaluating action programs: Readings in social action and education* (pp. 105-117). Boston: Allyn and Bacon.
- Beswick, R. (1990). *Evaluating educational programs*. Eugene, OR: ERIC Clearinghouse on Educational Management. (ERIC Digest Series Number EA54)
- Century, J. R. (1997). *The evaluator as technical assistant: A model for systemic reform support*. (Doctoral Dissertation, Boston University).
- Eash, M. (1985). Evaluation research and program evaluation: Retrospect and prospect. *Educational Evaluation and Policy Analysis*, 7 (3), 237-252.
- Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage Publications.
- Harari, D. (1974). *The role of the technical assistance expert*. Paris: Development Centre of the Organisation for Economic Co-operation and Development.
- Madaus, G., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. Jackson (Ed.), *Handbook of research on curriculum: A project of the American Educational Research Association* (p. 119-154). New York: Macmillan Publishing Company.
- McColskey, W., Parke, H., Harman, P., & Elliott, R. (1995). Evaluators as collaborators in science education reform. In R. O'Sullivan (Ed.), *Emerging roles of evaluation in science education reform* (pp. 71-89). San Francisco: Jossey-Bass.
- O'Sullivan, R. (1995). From judges to collaborators: Evaluators' roles in science curriculum reform. In R. O'Sullivan (Ed.), *Emerging roles of evaluation in science education reform* (pp. 19-29). San Francisco: Jossey-Bass.
- Shadish, W., Cook, T., & Leviton, L. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage Publications.

ASSESSING STUDENT OUTCOMES

Norma Dávila
University of Puerto Rico

Overview

Student academic achievement is often the main area of interest for educators and policy makers within any discussion of systemic educational reform. These discussions are usually centered on traditional test scores that may or may not reflect what is important for reformers and educators yet, for many, they are the only available mechanism to demonstrate the impact of an initiative. Finding and designing alternative ways to measure student academic achievement within the new parameters of systemic educational reforms has been a major challenge for both evaluators and reformers who have searched together for answers to accountability questions. This paper presents the evolution of, and the lessons learned from, a research approach to assessment of student outcomes, specifically of student academic achievement, being used by the Puerto Rico Statewide Systemic Initiative (SSI), which is one of the statewide systemic initiatives for science and mathematics sponsored by the National Science Foundation.

Definition of Outcomes and Outcome Variables

Weiss (1998) describes outcomes as "the end results of the program for the people it was intended to serve" (p.8) and further comments that outcomes are interchangeable with results and effects. Outcomes are certainly an end result of systemic educational reforms as well as of many other types of programs, but the nature and context of these initiatives requires a wider definition. For example, in systemic educational reforms, outcomes can be evident at the level of the classroom, school, district, or state. Evaluators of systemic educational reforms are usually interested in connections between different

interventions and outcomes, as well as in the factors that contributed to the occurrence of those outcomes.

Because of the additional dimensions of systemic educational reforms that differentiate these programs from other educational interventions, distinctions between outcome variables and outcomes need to be established. In systemic educational reform, an outcome variable is a quantity, dimension, or quality of the system subject to change because of the initiative. A systemic variable is an outcome variable that can be measured across the system such as student academic achievement in science and mathematics. In turn, an outcome for a systemic initiative is a change in an outcome variable directly attributable, or likely attributable, to the initiative, such as improvements in student learning as a result of participation in standards-based instruction in science and mathematics.

Importance of Student Achievement Outcomes within Systemic Educational Reforms

The central focus of most systemic educational reforms is the achievement of challenging academic standards that can be demonstrated through improvements in student academic achievement. Student academic achievement is interrelated with aspects of the initiatives such as their visions of quality education, expectations of performance for participants, definitions of equity, and designs of professional development interventions among others. Further, student academic achievement is a concrete indicator of progress that is associated with other areas of student success, such as college and job placement. Thus, systemic educational reforms are often expected to provide evidence of having an

impact on student academic achievement as an indicator of the value added by the reforms. Consequently, evaluators face the challenge of choosing an appropriate data collection and reporting design that meets the needs of the initiatives and of their multiple stakeholders.

The Evolution of a Research Approach in the Assessment of Student Outcomes

Just like many other systemic educational reforms in science and mathematics, the Puerto Rico SSI's central focus is the student as an active learner (Shields, March, & Adelman, 1998). The Puerto Rico SSI fosters the holistic development of the students in preparation for their participation in the next century as illustrated in the constructivist principles that guide this reform; the Puerto Rico SSI envisions the teaching and learning process as bi-directional and interactive with the guidance of teachers within the context of school environments (Davila, Vega, & Rodriguez, 1996). A participatory-research approach was selected for the evaluation and assessment of the Puerto Rico SSI, in general, and for the assessment of student academic achievement, in particular, because: (1) the philosophy that guides this initiative emphasizes participant empowerment and the development of self-sustaining communities of learners; (2) the size and scope of the initiative require the involvement of increasing numbers of individuals; (3) the Puerto Rico SSI's reformers and participants possess expertise in a diversity of areas that can significantly contribute to the successful implementation of such a model; and, (4) the literature available at the beginning of the initiative's implementation (in 1991) clearly demonstrated a need for results of systemic educational reform based on research (Davila, 1996).

Triangulation of results has been a major element of the design from the beginning of this reform. By comparing findings obtained using multiple quantitative and qualitative data collection strategies as suggested in the literature (Laguarda, Goldstein, Adelman & Zucker, 1998), the Puerto Rico SSI has

identified trends and made pertinent mid-course corrections within its encompassing systemic strategy. The Puerto Rico SSI's participatory research, evaluation, and assessment design involved all the different areas being addressed by this comprehensive science and mathematics reform (see Dávila & Gómez, 1994; 1995; Davila, Gómez, & Vega, 1996, among others, for specific examples). However, documenting and measuring student academic achievement was a major area of emphasis in this design because of (1) its importance within the larger context of systemic initiatives, and most importantly, (2) its value for the Puerto Rico SSI for decision-making purposes.

First Version of the Puerto Rico SSI's Model to Assess Student Academic Achievement

The first version of the model consisted of collecting and interpreting data at three different levels: (1) the classroom; (2) the initiative; and (3) the system (see Figure 1) (Puerto Rico Statewide Systemic Initiative, 1997). The description of each one of these levels follows. As part of their professional development, science and mathematics teachers learn to use authentic assessment strategies such as open-ended questions, performance tasks, portfolios, and multiple-choice questions that require higher order thinking skills to obtain information about student progress. Teachers use the results provided by these innovative strategies in their classrooms to (1) provide feedback to students about their performance and (2) modify their teaching, learning, and assessment practices. Teachers also translate these results into letter grades: schools provide grade distributions in terms of satisfactory (i.e., As, Bs, Cs) and unsatisfactory (i.e., Ds and Fs), before and after their participation in the Puerto Rico SSI, to identify trends in student academic achievement.

The initiative's staff developed a series of standards-based **pre/post** tests in science and mathematics to measure the value added by the systemic educational reform, as part of the

second level of the model. These tests included multiple-choice items that measure higher order thinking skills, open-ended questions, and performance tasks. Thus, assessment of student academic achievement was aligned at the classroom and initiative levels. Initially, all participating students took these assessments and, later, as the number of students and schools increased, representative samples of students were selected to represent their schools in the assessments.

The third level of the model consisted of external indicators of student progress for the overall K-12 system. The results of these tests provided other measures for the Puerto Rico SSI to "take the pulse" of the reform, even though they were not fully aligned with the standards-based reform. These indicators included an adaptation and translation of the National Assessment of Educational Progress (NAEP) that was administered in 1994 in both science and mathematics to samples of participating Puerto Rico SSI students (i.e., lower socio-economic levels), students from private schools (i.e., middle and upper middle socio-economic levels), and students from non-participating public schools (i.e., lower socio-economic levels). They also included other tests designed by testing corporations and administered by the Puerto Rico Department of Education, such as the **SENDA** and the Puerto Rican Competencies Test.

The first version of the model provided very useful information to the Puerto Rico SSI. However, as the needs of the initiative evolved, new ways to (1) look at student academic achievement; (2) provide specific formative feedback of student academic achievement to multiple players and stakeholders; and (3) design more mechanisms to drive the improvement of student learning in science and mathematics were imperative.

Second Version of the Puerto Rico SSI's Model to Assess Student Academic Achievement

The centerpiece of the second version of the model is the science and mathematics pre/post tests designed by the Puerto Rico

SSI's staff in an alliance with The College Entrance Examination Board (CEEB), which provided technical expertise for their administration and analysis (see Figure 2) (Puerto Rico Statewide Systemic Initiative, 1998). The new tests were designed to measure achievement gains over the course of one year, using publicly-released multiple-choice and open-ended items from NAEP (National Assessment of Education Progress) and TIMSS (Third International Mathematics and Science Study). The tests were administered to the fourth, eighth, and eleventh grades; all students from the 377 Puerto Rico SSI schools participated in this new assessment.

The new standards-based tests are scored using a scale equated with the TIMSS scale for item difficulty and student ability; a score of 500 in either scale equals the international average. By using a scale equivalent to that of the TIMSS scale, student scores can be compared with national and international benchmarks of student performance that allow the Puerto Rico SSI to place the progress of its students within the larger global context (see Figure 3). By sharing the schools' results of the pre-tests by content area with school principals and teachers, the school can assume responsibility for improving student learning that can be demonstrated in the post-tests. Further, the results of the pre/post-tests will be used to refocus the initiative's professional development activities, based on the content needs of the students.

Another key element of the second version of the model is the teachers' participation in parallel assessments; its main purpose is to identify and correct teachers' weaknesses in content. Teachers receive sets of items not included in the tests administered to their students (but similar in approach and content) during a professional development session and are asked to respond to them anonymously. An item by item analysis of the distribution of their responses leads to a discussion of common misconceptions and of ways to correct them. The information provided by these analyses provides another mechanism for refocussing the initiative's

professional development activities to address specific content needs of the teachers.

An external criterion now included in the Puerto Rico SSI's assessment of student academic achievement is the results of the college admissions tests administered by the CEEB. Since equating studies between the SAT and the CEEB mathematics tests show a correlation of 0.87, the Puerto Rico SSI can confidently compare the results of students in the mathematics test of the CEEB with those of mainland students in the mathematics test of the SAT (see Figure 3).

Another external criterion is the ratios of college admissions to the University of Puerto Rico System, which is the most competitive university system of the Island. College admissions ratios of Puerto Rico SSI participants are being analyzed by length of initiative intervention (i.e., intermediate school only vs. intermediate and high school). Distributions of chosen field of studies upon admissions are being analyzed in a similar way.

The evolution of the first and second versions of the Puerto Rico SSI's model to assess student academic achievement as an outcome of systemic educational reform show that considerable organizational learning has taken place within the Puerto Rico SSI. The following section addresses some of the lessons that the leadership of this reform has learned in the process of designing these models.

Lessons Learned

The process of designing the two versions of the assessment model required intense reflection and thinking by the leadership of the Puerto Rico SSI at multiple levels. Since the first version of the model had provided the initiative with very useful information over the years, it was difficult at first to make the decision to find another way to measure student academic achievement. However, the national exposure and dissemination of the TIMSS reports since 1997 was certainly a factor that prompted us to look for other alternatives more in tune with the evolving needs of the reform. Using publicly-released

items from NAEP and TIMSS represented a major cost-saving step, since the items had already been developed; but, without the vision and expertise of The College Entrance Examination Board (CEEB), we would not have achieved the same results. At the same time, the Puerto Rico SSI staff is influencing the test design vision of this major player in education by emphasizing and **modelling** the use of national standards to guide test design. Further, the involvement and engagement of teachers in the professional development exercises described above gave us pleasant surprises, since they sincerely enjoy the experience of looking at their own performance and, most importantly, they grow professionally and personally in the process.

Final Comments and Next Steps

One of the major challenges currently faced by evaluators and reformers who work with systemic educational reforms is the need for common metrics of student academic achievement. This is a recurrent theme in meetings sponsored by the National Science Foundation and it is evidently a high priority on the national educational reform agenda. We believe that the models presented in this paper can contribute to advance the design of such metrics.

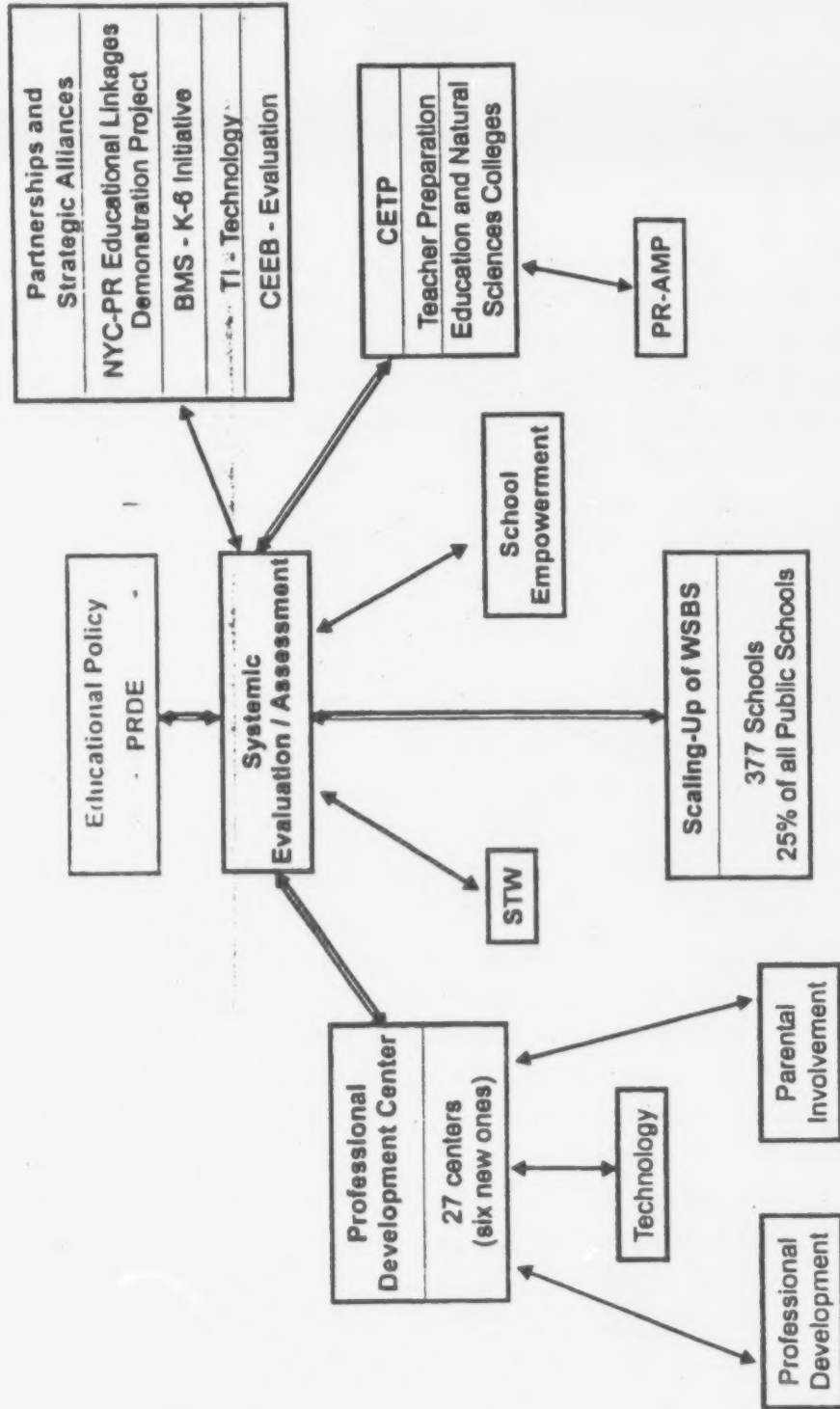
References

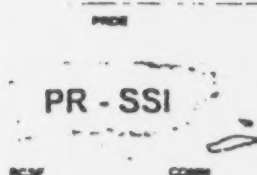
- Davila, N. (1996). *Design and use of an interactive evaluation model to assess systemic change in science and mathematics education*. Paper presented at the ASCD 51st. Annual Conference, New Orleans.
- Davila, N., & Gómez, M. (1994). *Assessment of the impact of a new curriculum on systemic change*. Paper presented at the AERA Annual Meeting, New Orleans. ERIC DOCUMENT 390 307
- Davila, N., & Gómez, M. (1995). *Evaluation of school-based regional dissemination centers as scale-up mechanisms for systemic educational reform in science and mathematics*. Paper presented at the

- AERA Annual Meeting, San Francisco.
ERIC Document 310 702.
- Davila, N., Gómez, M., & Vega, I.Y. (1996). *Evaluating the transformation of the teaching/learning culture of schools involved in systemic science and mathematics educational reform*. Paper presented at the AERA Annual Meeting, New York. ERIC Document 395 803
- Davila, N., Vega, I.Y., & Rodriguez, J. (1996). Fomentando el desarrollo integral de los y las estudiantes mediante un currículo constructivista-interaccionista de ciencias y matemáticas: **Implantación y resultados**. [Fostering student holistic development through constructivist-interactionist science and mathematics curricula: Implementation and results.] Paper presented at the Eighth National Education and Thought Encounter, Ponce, Puerto Rico.
- Laguarda, K.G., Goldstein, D.S., Adelman, N.E., & Zucker, A.A. (1998). *Evaluation of the National Science Foundation's Statewide Systemic Initiative (SSI) Program. Assessing the SSI's impact on student achievement: An imperfect science*. Menlo Park, CA: SRI International.
- Puerto Rico Statewide Systemic Initiative. (1997). *Phase II Proposal*. San Juan, PR. Author.
- Puerto Rico Statewide Systemic Initiative. (1998). *Performance Effectiveness Review Report*, December 1998. San Juan, PR. Author.
- Shields, P.M., Marsh, J.A., Adelman, N.E. (1998). *Evaluation of NSF's Statewide Systemic Initiatives (SSI) Program: The SSI's impact on classroom practice*. Menlo Park, CA: SRI International.
- Weiss, C.H. (1998). *Evaluation: Methods for studying programs and policies*. Upper Saddle River, NJ: Prentice Hall.

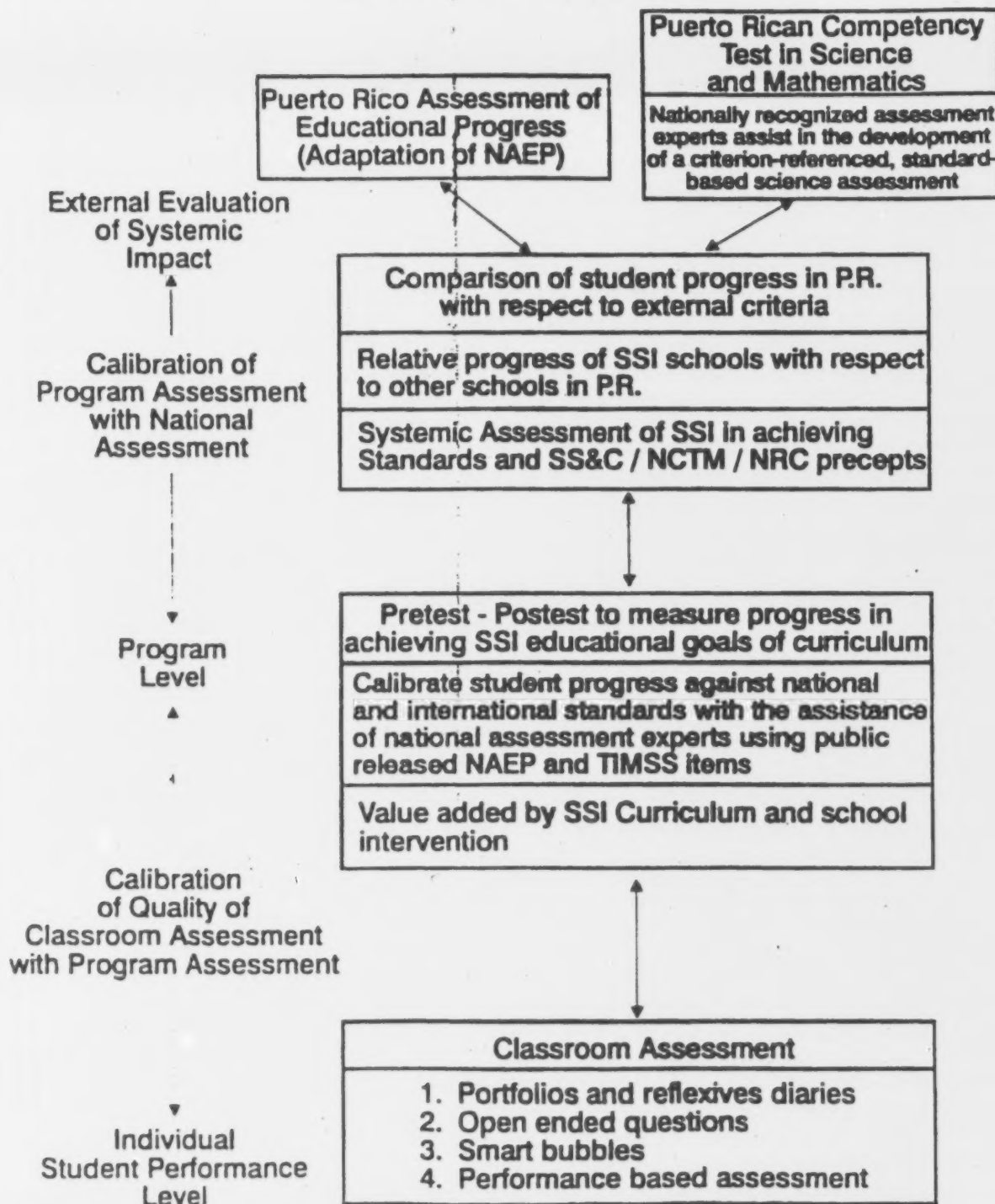


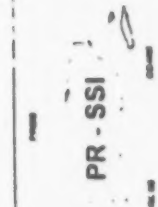
PR-SSI Phase II Evolving Implementation Plan





Systemic Assessment of Student Performance Value Added by the Reform Three Tier Approach





**Differences in Performance in TIMSS / NAEP
Pre-Test between PR-SSI School Generations
*Benchmarked Against International TIMSS Standards**

Beginning Eighth Grade Student Performance in Science Multiple-Choice Tests

Population	Mean Score	International Benchmarks: Seventh Grade Performance Mean Score			
		Country	Score	Country	Score
All Eighth Graders	456	Romania	452	Iceland	462
Generation 1	472	Scotland	468	Spain	477
Generations 2-3	455	Romania	452	Iceland	462

Beginning Eighth Grade Student Performance in Mathematics Multiple-Choice Tests

Population	Mean Score	International Benchmarks: Seventh Grade Performance Mean Score			
		Country	Score	Country	Score
All Eighth Graders	425	Portugal	423	Lithuania	428
Generation 1	444	Greece	440	Cyprus	446
Generations 2-3	423	Portugal	423	Lithuania	428

* The CEEB equated the TIMSS / NAEP scale with the TIMSS International scale to permit valid comparisons.



Evidence of Differential Impact on Students Exposed to a Continuum of Reformed Teaching/Learning Experiences vs. Shorter and Discontinuous experiences

[TIMSS/NAEP scale equated to international levels of performance in TIMSS]

Comparison of PR-SSI Beginning Eleventh Grades Performance in Science and Mathematics

Subject Area	Population	Mean Score	International Benchmarks: Eleventh Grade Mean Score			
			Country	Score	Country	Score
Science	All PR-SSI Students	545	Norway	544	Iceland	549
	Schools with Students who had four years of PR-SSI experience	584	Sweden	559	---	---
Mathematics	All PR-SSI Students	419	South Africa	356	Cyprus	446
	Schools with students who had four years of PR-SSI experience	451	Cyprus	446	United States	461

Scores of PR-SSI TIMSS based test were equated to international scales by the CEEB within an error of 8 points

The performance of all eleventh grade students is compared against that of students from high schools whose eleventh grade students have four years of PR-SSI experience

RESUMEN DE RESULTADOS DE LA PREPRUEBA
REGIÓN HUMACAO - CUARTO GRADO

CODIFICACIÓN: 33308
ESCUELA: RUFINO VIGO
PUEBLO: HUMACAO

PRSSI

PUERTO RICO STATEWIDE SYSTEMIC INITIATIVE

SELECCIÓN MÚLTIPLE

Ciencias		
Puntuaciones	Núm. Estudiantes	Porcentila
900	0	-
850 - 899	0	-
800 - 849	0	-
750 - 799	0	-
700 - 749	0	-
650 - 699	1	99
600 - 649	3	98
550 - 599	15	90
500 - 549	29	71
450 - 499	28	46
400 - 449	22	24
350 - 399	8	11
300 - 349	7	4
250 - 299	0	-
200 - 249	1	1
150 - 199	0	-
100 - 149	0	-

N 114
Mediana 258
Promedio 478
Desv. Estándar 77
Máximo 674
Mínimo 203
Omitidos 1

Matemáticas		
Puntuaciones	Núm. Estudiantes	Porcentila
900	0	-
850 - 899	0	-
800 - 849	0	-
750 - 799	0	-
700 - 749	0	-
650 - 699	0	-
600 - 649	0	-
550 - 599	1	99
500 - 549	12	94
450 - 499	23	78
400 - 449	30	55
350 - 399	25	31
300 - 349	16	12
250 - 299	6	3
200 - 249	0	-
150 - 199	0	-
100 - 149	0	-

N 113
Mediana 264
Promedio 414
Desv. Estándar 68
Máximo 590
Mínimo 264
Omitidos 1

PREGUNTAS ABIERTAS

Ciencias		
Puntuaciones	Núm. Estudiantes	Porcentila
900	0	-
850 - 899	0	-
800 - 849	0	-
750 - 799	0	-
700 - 749	0	-
650 - 699	1	99
600 - 649	2	98
550 - 599	9	93
500 - 549	18	82
450 - 499	28	62
400 - 449	23	40
350 - 399	23	20
300 - 349	8	6
250 - 299	3	1
200 - 249	0	-
150 - 199	0	-
100 - 149	0	-

N 115
Mediana 276
Promedio 447
Desv. Estándar 79
Máximo 683
Mínimo 253
Omitidos 0

Matemáticas		
Puntuaciones	Núm. Estudiantes	Porcentila
900	0	-
850 - 899	0	-
800 - 849	0	-
750 - 799	0	-
700 - 749	0	-
650 - 699	0	-
600 - 649	3	99
550 - 599	6	95
500 - 549	10	88
450 - 499	20	75
400 - 449	22	56
350 - 399	27	35
300 - 349	21	14
250 - 299	3	3
200 - 249	2	1
150 - 199	0	-
100 - 149	0	-

N 114
Mediana 244
Promedio 419
Desv. Estándar 83
Máximo 622
Mínimo 223
Omitidos 0

PR-SSI Project - September 1998

Multiple Choice Science scores by school sorted by mean - 11th grade

Summaries of SCORE Multiple Choice Items				
By Levels of SCHOOL				
Variable	Value Label	Mean	Std Dev	Cases
For Entire Population		544.641	81.086	906
SCHOOL	61531 University Gardens	642.015	76.026	88
SCHOOL	27318 Francisco García Boyrié	593.179	64.288	50
SCHOOL	20560 Pablo Colón Berdecia	550.900	79.701	156
SCHOOL	26021 Luis Muñoz Marín	542.252	71.697	195
SCHOOL	66209 Luz A. Calderón	542.138	68.414	135
SCHOOL	75739 Rosalina C. Martínez	517.721	68.165	37
SCHOOL	51698 Luis Lloréns Torres	514.415	47.567	64
SCHOOL	34256 Casiano Cepeda	505.420	76.274	46
SCHOOL	31716 Juan J. Mauney	496.985	69.185	135
Total Cases = 906				

PR-SSI Project - September 1998					
Multiple Choice Math scores by school sorted by mean - 11th grade					
Summaries of SCORE Multiple Choice Items					
By Levels of SCHOOL					
Variable	Value Label	Mean	Std Dev	Cases	
For Entire Population		419.303	78.377	900	
SCHOOL	61531 University Gardens	537.846	81.779	89	
SCHOOL	20560 Pablo Colón Berdecia	426.271	71.143	238	
SCHOOL	66209 Luz A. Calderón	423.176	57.005	71	
SCHOOL	26021 Luis Muñoz Marín	413.475	60.809	175	
SCHOOL	75739 Rosalina C. Martínez	406.326	52.159	38	
SCHOOL	27318 Francisco García Boyrié	387.695	61.789	65	
SCHOOL	31716 Juan J. Maunex	381.919	55.334	135	
SCHOOL	51698 Luis Lloréns Torres	375.830	68.332	89	
Total Cases = 900					

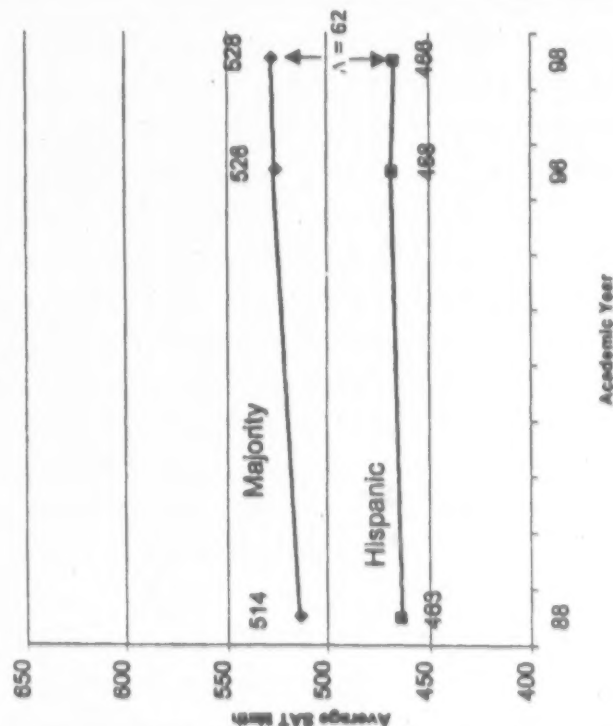
PR-SSI Project - September 1998					
Multiple Choice Math scores by school sorted by mean - 11th grade					
Summaries of SCORE Multiple Choice Items					
By Levels of SCHOOL					
Variable	Value Label	Mean	Std Dev	Cases	
For Entire Population		419.303	78.377	900	
SCHOOL	61531 University Gardens	537.846	81.779	89	
SCHOOL	20560 Pablo Colón Berdecia	426.271	71.143	238	
SCHOOL	66209 Luz A. Calderón	423.176	57.005	71	
SCHOOL	26021 Luis Muñoz Marín	413.475	60.809	175	
SCHOOL	75739 Rosalina C. Martínez	406.326	52.159	38	
SCHOOL	27318 Francisco García Boyrié	387.695	61.789	65	
SCHOOL	31716 Juan J. Mauné	381.919	55.334	135	
SCHOOL	51698 Luis Llorens Torres	375.830	68.332	89	
Total Cases = 900					



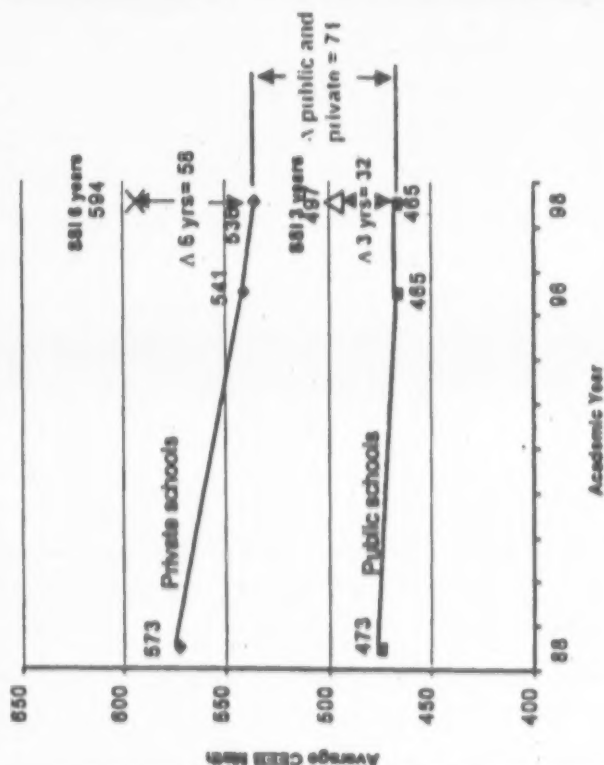
Longitudinal Study of Performance of Former PR-SSI Participants: Scores in College Admissions Tests (CEEb test and SAT test)

[Evidence of differences between continuous and interrupted reform interventions on student performance]

SAT Mathematics
Hispanic compared to Majority population
(Mainland)



College Board - Mathematics
(CEEb equating studies between CEEb and SAT tests show a correlation of 0.87 between both tests)



Private schools: students from families with middle and upper middle income backgrounds

Public schools: students from families with lower income backgrounds

UNDERSTANDING THE VALUE OF NSF'S INVESTMENTS IN SYSTEMIC REFORM

Mark St. John
Inverness Research

The Challenge

I was invited by a sub-committee of Congress to talk to its members about the value of the National Science Foundation (NSF) investments that are being made in the Urban, State, and Rural Systemic Initiatives. The questions they asked are very simple, very basic: "Are they working?" and "What are the benefits [in return] for the millions of dollars being invested?" Very simple and reasonable questions to ask-and yet very difficult to answer. My first question as an evaluator was: "How do you even think about that question? What are the kinds of benefits that one might reasonably expect to come out of the investment made in the Systemic Initiatives?" How do you conceptualize that kind of return on investment?

So today I want to talk to you-and think out loud with you-about that question. As my friend, Patrick Shields, would say, when you think about evaluating systemic reforms, the answer to the question depends very much on what you mean by "evaluation," and it depends very much on what you mean by "systemic reform." So let me focus briefly on what I mean by each of those terms.

Evaluation, in this case, refers to figuring out the value, the contribution, and the benefits that accrue from the public investment that is being made in the Systemic Initiatives. Today, I speak primarily from the perspective of an evaluator, a professional evaluator, who is trying to think about these issues in a substantive, scientific way. I am thinking primarily as an evaluator-and not as a science educator. That is, I am thinking about the serious problem of assigning value to an investment that is being made according to the "systemic theory" of improving science education. I am not thinking as a science educator trying to use evaluation as a way to

further the cause of science education. Thus, I am trying to think about the issue of evaluation objectively and scientifically, and not politically. That is a different exercise. I am making an effort to identify the real value of these investments-not those values that may or may not have political currency. I was trained as a physicist. So I began to think about this as a physicist. What do I observe in the field, and in my interviews, that would qualify as a significant contribution and value? What might we infer from the evidence about the nature of the value and benefits that come from NSF's work? What are the kinds of things we could really say about NSF's work if there were no political pressures upon us? How would we think about evaluation and systemic reform in a clear and straightforward fashion?

The Nature of Systemic Change

Now for part two of defining my terms. When I refer to "systemic reform," I have a fairly simple idea in mind. I think systemic reform is quite right in its analytical insight. The theory of systemic reform says: The instruction that students receive-the quality of their learning experiences in schools-is directly shaped by the system, by the political and institutional context, that surrounds the classrooms they are in. The theory of systemic reform says: Any attempt to improve the quality of instruction that students receive must, to be successful, assume a systemic perspective in the design and implementation of that effort. For example, according to the theory of systemic reform, you cannot just pick one element of the system (for example, curriculum) and work on it, and then expect instruction to improve. The system that shapes instruction is itself composed of multiple, hierarchical, interacting, and complex

systems-systems of professional development, assessment, curriculum, school governance, and others. We know that good instruction happens when there is a convergence, and alignment, of many necessary, but not sufficient, sub-components of the larger system. That is, improvement in each system component, such as curriculum, professional development, and assessment, is a necessary, but not sufficient improvement when it comes to increasing the quality of instruction and raising the level of student achievement.

An analogy, and friends who know me are tired of this example: But, in the airline system, it is quite clear there are many necessary, but not sufficient, components that come together to give us a safe and reliable airline system. It is not enough to have good pilot training. Good pilot training is clearly necessary. But it is not sufficient unto itself. One also needs well-designed, well-maintained airplanes, good air traffic control, and good airports. All of these system ingredients are simultaneously required for safe airline operation. It may well be possible to greatly improve the quality of pilot training and yet still have an unsafe airline. Also, we cannot improve professional development this year and say that we will handle airport design and safety next year. All of the components must be simultaneously present-working together and of high quality. The real issue in all of this is that each of these components may well be necessary, but they are not sufficient. You cannot get seven out of eight right, and leave the eighth undone. Think about the airline system as a helpful guide in understanding how a systemic approach and perspective are necessary in effecting educational change.

So, in education, as in other complex endeavors, there are many necessary yet not sufficient system supports that must be present if the system is to function well. That is one important notion that underlies the theory of systemic reform.

Secondly, there is another aspect of systemic reform that is equally important—but is often overlooked. That notion is that the people who do the work of improving the

system must be those living and working within the system. That is, in a systemic reform effort, the changer is the changee. It does not help very much to have external agents (such as university professors, no offense) do a lot of research in their laboratories, and then present the results to school districts. It is the people who live within the system (those who work at the state, district, and school levels) who must gain the capacity and have the resources to improve the functioning of their own systems. That means that they must have the capacity and expertise to bring about intelligent change. Further, this capacity must consist of internal skills and knowledge, as well as access to external resources and expertise. So we see that "capacity building" is also a very important part of the theory and work of systemic change. Because the changer must be the changee, the real work of systemic reform lies in building the capacity of those who are key change agents within the system. Finally, this idea has re-implications for helping us think about how we might evaluate NSF's investment in systemic **change**—which, as you will remember, was our original charge.

Issues in Evaluation

As we look more carefully at the role of evaluation, let me share with you some of the concerns I have. There are many definitions of "evaluation," but I want to propose what I consider a very simple one: *I think evaluation is about helping people understand more clearly, and in more powerful ways, what is actually happening.*

Evaluation should help clarify what is real: When, as evaluators, we study a program or a curriculum, we should help to conceptualize, explain, and illuminate what is actually happening. That is a different process from helping NSF "make its case." It is a little different from helping a state systemic initiative make its case. In our evaluations, we should be careful about the inferences that people are encouraged to draw from our work. We should be encouraging caution, and we should be actively discouraging false

inferences. We should be dispelling-and not **creating**—myths and overly simplistic conclusions.

Evaluation should also help educate and inform the public about the realities and complexities of education. As evaluators, we should use our insights to help people think more powerfully about the way the system works, about the way learning happens. We should be educative. And we evaluators, who have a lot of experience in seeing how the system does and does not work, should share our insights with others. The power of these insights is that they are grounded in, and intertwined with, the data that we gather. In a way, you could say evaluation should be about identifying and reducing any gaps that may arise between rhetoric and reality, between mythology and actuality. That is what I think evaluation should be doing.

In physics education, there are frequently discussions about student **misconceptions**—about the ideas students have about the way things work that are not at all congruent with reality. At this point, I want to share with you certain evaluation misconceptions, or perhaps I should call them “accountability misconceptions,” that people have about education that are not congruent with reality.

At the risk of seeming naive: I know there are “political realities” that help to generate and perpetuate these kinds of misconceptions. But I nonetheless think it is in our long term interest to be critical of misconceptions and false inferences that try to wrap themselves in the respectable cloak of evaluation and accountability language.

There is one misconception, for example, that is fairly prevalent. I call it: “the last input is the only input.” That is, it is often argued that we can evaluate the quality of a fourth grade teacher by assessing the achievement of the students of that fourth grade teacher, say in mathematics. The fact that students have only been in fourth grade for four months and have spent eight years in other classrooms or settings—well, we do not address that reality. We create an accountability and evaluation system that responds as if the fourth grade teacher were solely responsible for what those

students know and are able to do in mathematics.

Another misconception of the same ilk: We say, if only teacher preparation programs were better, then teachers would be of good quality. In California, teacher preparation is a year-and-a-half program. With little control over who enters the program, and relying on sixteen previous years of schooling in the disciplines, teacher preparation programs are somehow held solely responsible for the quality of beginning teachers. This kind of over-simplistic assignment of cause may be politically attractive, but it is scientifically very, very weak.

A further, and closely related, misconception equates the absolute value of something to the value-added component. For example, there are many programs that are supposed to have an incremental effect on something—but those programs should be judged by the value they add, not by the absolute value of their products. Let me be concrete here. You do not judge the quality of a psychiatrist by the absolute level of the sanity of his or her patients. No. The job of psychiatrists is to make their patients saner than they would otherwise be. If absolute value of sanity were the criterion, it would be easy to be a good psychiatrist: you would simply start with the sanest people. So it is important to make the point that a good **psychiatrist** contributes a greater sanity to the patients: He or she adds value, makes them slightly saner than they would have been otherwise.

In the same sense, you should not say that a “good? school is a school where the students all have high test scores. It is exactly equivalent. Good schools add significant value to the knowledge and skills that those students bring to the school. I have been in some very prestigious schools that, I would argue, add very little value to the bright kids who come in. On the other hand, I have been, just recently, in some wonderful schools where the students are scoring in the lowest twentieth percentile—schools that are, I would argue, wonderful schools because their skillful, dedicated teachers are bringing very disadvantaged students up from zero to

twenty percent. Yet, none of that is acknowledged in a simplistic accountability system. In fact, our very language is confusing here when we talk about "high performing and low performing schools"—as if the school itself were taking the tests. No, we should be focusing on schools at which students, on average, are performing or not performing, well on tests, because the performance of the school, as a provider of instruction, is not at all the same as the performance of the students.

Thus, my impassioned plea here is to evaluators, to suggest that it is our collective responsibility to address **these** kinds of misconceptions and over-simplistic notions and, in doing so, **to help hold accountability accountable**. It is ironic that accountability itself—the whole movement of creating standards, tests, and high stakes accountability systems—gets a free ride. Where is the evidence that accountability systems, used as they are now being used, increase student learning? I would assert that it is not clear at all that accountability exerts a positive force on the system, so I argue that we need to be critics of unexamined and even incorrect notions. We need to be pushing more sensible, less political interpretations of how value is assigned to schools, to programs, and to systemic initiatives.

Now, to consider one final type of misconception that is attractive because it makes the world simpler to understand. The only problem is that it is wrong. (Some philosopher once said, "To every complex problem, there is a simple-and incorrect—answer.") This misconception lies in the area of "attribution." I would call it **the single variable, or sole-agent, problem**. The assumption that evaluations of the NSF and other reform efforts seem to make sometimes is that the program being studied is the only one in town, the only one doing anything. It is as if everything else is static and that we are working in a laboratory setting where all the other inputs are absent and all the variables held constant.

But that is far from the truth. In the very language that is used, you often hear confusion, not only about the fact that NSF is

not a sole agent, but sometimes it is not an agent at all. We hear that "NSF is doing systemic reform" in a given city or state. But NSF does not itself "do" systemic reform. NSF funds people and programs under a theory of systemic reform—people and programs involved in a variety of activities. Also, while such funding may actually represent a lot of money for NSF, it may not be much money for the system in which it is working. Some reform activities funded by NSF can be one of many different things that are happening in a larger state and district reform context. If you go to big districts and ask about what is happening there, it might be a long time before anybody mentions NSF. NSF money, and NSF activities, are inevitably a piece, a small part, of the systems they are trying to influence. The NSF activities are often not "at the center of the state's or district's radar screen." So the image of a state or district that is engaged in a systemic reform effort funded by NSF is often seen from what might be called an "NSF-centric" point of view.

It is important to note here that NSF-funded activities are NOT insignificant. They may well be doing important work within the local setting. But they are not the whole piece, and very often the NSF work is irresolvably mixed in with many, many other activities and reform efforts.

Also, in its Systemic Initiatives, NSF is working a great distance from the classroom and the students. That is both its power and its weakness. As a metaphor, imagine that NSF funds and activities are like shining a light down a cone. In this picture, NSF-funded activities are strategically designed to influence state and district reform activities. And they get leverage because the state and district reform activities cover a larger cross-sectional area than the NSF activities. Further, those state- and district-level reform activities, we hope, build long-term state and district capacities—so that those states and districts can, in a sustainable way, create better mathematics and science programs. In turn, these programs can help to improve science, mathematics, engineering, and technology instruction on a broader scale. Finally, such

instruction improves student achievement on an even greater scale.

The problem is that as we shine the light down the cone, we encounter the inverse square law. The light gets inversely dimmer as the square of the distance down the cone increases, so by the time you get to the end of the tunnel-to the cross section where student achievement lies-the light is very diffused, and very dim.

To make matters more difficult, this second cone on the graphic may represent another reform effort. It might, for example, be school restructuring. It might be an accelerated schools program. It might be reduced class size. So, at any given time, in any given district, five, six, ten, twenty, forty other lights may be shining down this cone. And these lights converge in very non-linear ways-so that the final illumination of the cone at the level of student achievement is very mixed, noisy, and undifferentiated in its mixture. In evaluation language, you have got some real attribution problems here. The more distant you are, the more attribution problems you have.

Let me summarize some of the real difficulties we have in establishing the value of NSF's investments in systemic reform. First, we have problems with the scale of the investment. In some of the SSIs, the NSF contribution is on the order of a couple of dollars per student per year. This is a low-level investment if what we envision is direct impact at the student level. (Let me add, for the record, that there is a paradox in this business. It is very hard to find the results of your work in improved student test scores. But if you are concerned with professional development, for example, you need to be focused on student learning and student achievement. So it does not mean that you ignore the details of how students learn and how they are actually doing. You remain highly focused on it. However, it does not work the other way: You cannot operate in the reverse. You cannot use student achievement as a measure for assessing the success or effectiveness of professional development. The reason for this, as we have shown, is that there are many streams that lead into the lake

of successful student achievement. You have to think about this to recognize that there is quite a paradox here.)

Another analogy will reinforce this very important point. Let us presume that you had a reform view of agriculture and that you wanted to help increase the use of soy protein and decrease the dependence on beef for protein. So you might devise programs that help farmers improve the quality of their soil so that it is better for growing protein. If this program is successful, you will be able to provide a professional development plan to farmers, teaching them how to improve the quality of their soil. And, it would be important for them to have a vision of soy production as their ultimate goal. But it would be a mistake to evaluate the professional development aspects of the program by measuring the degree to which soy production is actually increased-especially in the short term. Why? Because there are many components that are necessary but not sufficient for creating an increase in soy production. One is the ability to develop appropriate soil. But one must also have the necessary amount of sun, water, and seed. Even more important, perhaps: There must be the demand for soy and the marketplace economics that makes it profitable to shift from beef to soy. So, we see that it is very possible to have an excellent professional development program for farmers and yet realize very little immediate increase in the production of soy. I think an equivalent argument holds for the evaluation of teacher professional development programs on the basis of student achievement.

Finally, there is the issue of the scale of the investment. Many NSF investments are very small compared to the scale of the system on which they are intended to impact. There is also a serious question about the time scale of the investment. When is it evident that systemic reforms are paying off? After two months, six months, one year, five years? Because investment in systemic change is about building capacity, the ultimate pay-offs are often delayed and very diffuse. Further, what is happening while one is waiting for the "downstream benefits" of the "upstream

investment" that NSF systemic investments represent? Remember that there are many, many other things happening simultaneously. It is a fact that nothing is standing still while all of this is going on: there are multiple, ever-changing sources of reform. So, all in all, at the classroom and student level, you might say that there is a very low signal-to-noise ratio—all of which makes it difficult and probably fundamentally impossible to evaluate NSF's systemic investments by directly connecting them to increases in student achievement.

The Real Benefits of NSF Investments in Systemic Reform

When we did a study for the National Academy of Sciences, and when we evaluated the data we had gathered from many states and districts engaged in systemic reform, some things began to emerge for us. We began to sense the probability that a particular NSF project would succeed. But only a probability. Because the success of any individual reform effort depended heavily on factors that were out of the control of those involved in promoting the reform, we were actually not very good at predicting success.

The equation presented here illustrates our current thinking about some of the most important things determining the probability that a reform effort will have real impact on the system it is seeking to improve. That probability depends, we would argue, on some factors with which you are all familiar. (These are ideas and, factors that you reformers and evaluators work with intuitively every day.) In the equation, the basic idea is that success depends upon the capacity that exists within the system, on the demand for the reform that is being promoted, and it is inversely proportional to the system barriers that exist.

The *L* in the equation stands for Leadership. When we visit states or districts, it becomes clear that the single most important factor is the quality, expertise, commitment, and political power of the leadership that is promoting the particular reform effort. Who is there to do the work?

Are they skilled leaders? Look at the successful initiatives we have seen over time and the one common factor that unites them is the presence of an individual, or a core of individuals, who are highly skilled, both at the district and state level.

D stands for Design. Design here refers to the knowledge and expertise that exist within the reform itself. The probability of a systemic reform succeeding is greatly enhanced by the presence of well-designed curriculum, well-designed assessments, and well-designed standards. There is also an element of design and sophistication that one looks for in the way in which a reform initiative is planned and the way in which its planners conceptualize and promote an overall "change strategy."

PRI stands for Policy and Reform Infrastructure. The question here is: Does this state, this district, have ways of doing work that involve system-wide changes and improvements? I was in an Appalachian district recently that had no such reform infrastructure at all—no concept of how to create district-wide change or improvement. The administrators were struggling simply to operate their district. They had no means of working to improve something. It was a foreign notion.

\$ represents Discretionary Dollars: Dollars that are allocated specifically for reforms in mathematics and science also increase the probability of success.

So these factors, which are internal, are all about the capacity of the state or district to launch and implement a systemic and system-wide reform. But to be successful, it is necessary to have more than these internal capacities. Political and public demand for the reform ideas being promoted is also necessary. To mount a successful systemic initiative, you need both capacity and demand. You can have all of the capacity in the world, but if there is no demand, then you only have a "supply side" reform. Or, you can have demand for change and an external commitment, but no capacity within the system to provide it. You need both.

Further, you face barriers that have nothing to do with mathematics and science

reform. In fact, most of the factors that determine the ultimate impact of mathematics and science reform do not come out of the activities conducted in the name of mathematics and science reform, but are a product of larger forces. These forces are detailed in the denominator of our equation. For example, the scale (*S*) of the system plays a large role in shaping the probability of success. Big states are much harder to change than small states. Big districts are harder to work with than small districts. This is because large systems tend to become fragmented, and the fragments are often at war with each other. Thus, it is hard to get a coherent reform underway. Also, in large systems, inertia is simply a greater problem.

There are several other system issues and events that greatly affect the progress and ultimate impact of NSF's systemic reforms. For example, states and districts may be swept by political "cross currents" (*P*), with progressive and conservative points of view battling for control. Or, the district may be struggling with severe financial (*F*) problems. Or, generally, the state or district may be experiencing great instability (rapid changes in leadership or structure or vision), so there is a certain amount of "turbulence" (*T*) buffeting the system.

And, ironically, one of the greatest threats to reform is the existence of other reforms. Many systems are currently suffering from what might be called "reform overload" (*RO*). Mathematics and science reforms increasingly have to compete with other priorities (*CP*)—for example, literacy.

So, what does this tell us? It suggests that there are areas within the system that NSF can influence and others that it cannot. It thus becomes clear that NSF is not in a position to DO the work of reform. At best, I would argue that the role of the National Science Foundation, the role of federal investment, is to *increase the probability* that a state or district will continue to improve its own mathematics and science programs.

I suggest that the appropriate measure of the value of NSF's investments in systemic reform lies in assessing the degree to which those investments have increased the enduring

capacity of states, districts, and schools to design, initiate, and sustain high quality improvement efforts in mathematics, science, and technology.

To respond to members of Congress who ask me how we should evaluate the NSF's investments, I would say the following: When NSF funds a five-year systemic initiative, and when the NSF-funded work has been completed, then that state or district should have enhanced capacity to continue the improvement of its mathematics and science programs. This capacity I am speaking of is not abstract: It consists of those factors that are listed in the numerator of our equation. For example, systemic initiatives should definitely leave in place leaders who are highly knowledgeable and skilled in mathematics and science reform. The vision of what good mathematics and good science is should be more sophisticated as a result of the initiative. Finally, there should also be an increase in the degree to which the state or district addresses curriculum and professional development issues with well-designed programs. The state or district should, in the long run, be better connected to multiple resources and multiple sources of expertise.

Thus, NSF's investments can and should be evaluated by assessing the degree to which they build the state or district capacity for initiating, and sustaining, further reform. NSF should not, I would argue, be held accountable for what the state or district does with that capacity. That is not the federal role, and it is beyond the control of NSF. But I do think that NSF should be held accountable for the capacity building that is a central goal of systemic reform. Measures of capacity are not impossible to design, and evaluators, I would argue, are capable of assessing the degree to which the capacity of a system is, or is not, increased over several years. By putting the emphasis on capacity, it could help to sharpen the focus of NSF's systemic initiatives. By contrast, if we insist on assessing the value of NSF's work by assessing student outcomes, then such evaluations will do much to confuse all involved and, ultimately, blunt the effectiveness of the work of their initiative.

In conclusion, I would add one final thought. Even in terms of politics and the political considerations that are always a factor, there is a fundamental wisdom in focusing on the capacity- building nature of these initiatives. I would point out that it is a very dangerous political strategy to design evaluations to serve short-term political goals. To argue that NSF's systemic initiatives are in themselves directly causing increased student achievement in the short term is to stretch the

trust and credibility that people are willing to give to evaluation claims. I think it is a wiser long-term political strategy to document the contributions the systemic initiatives are, in fact, making. It is a better long-term strategy to tell the truth about what these initiatives can and cannot do, and then document well their real benefits to the states and districts they are serving. That is, for me, a more grounded and satisfying way to go.

THE EVALUATION OF NSF'S INVESTMENTS IN
SYSTEMIC REFORM INITIATIVES

AN IMPORTANT ROLE FOR EVALUATION

TO HELP PEOPLE MORE CLEARLY UNDERSTAND WHAT
IS ACTUALLY HAPPENING

A N D

TO ENHANCE THE WAYS IN WHICH THEY ARE ABLE TO
THINK ABOUT WHAT IS HAPPENING

(TO REDUCE THE GAP BETWEEN REALITY AND
RHETORIC)

THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES

MISCONCEPTIONS AND MISLEADING INFERENCES

THE LAST INPUT IS THE ONLY INPUT

**EVALUATION OF FOURTH GRADE MATH TEACHER
TEACHER PREP AND TEACHER QUALITY**

ABSOLUTE VALUE = VALUE-ADDED

(Psychiatrist are evaluated by the sanity of their patients.)

SCHOOL QUALITY AND STUDENT TEST SCORES

SOLE AGENT THEORY

. NSF DOES SYSTEMIC REFORM... .

THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES

KEY ASPECTS OF THE THEORY OF SYSTEMIC REFORM

***CLASSROOM PRACTICES AND STUDENT LEARNING EXPERIENCES ARE SHAPED BY THE SYSTEM OF EDUCATION THAT SURROUNDS THEM.**

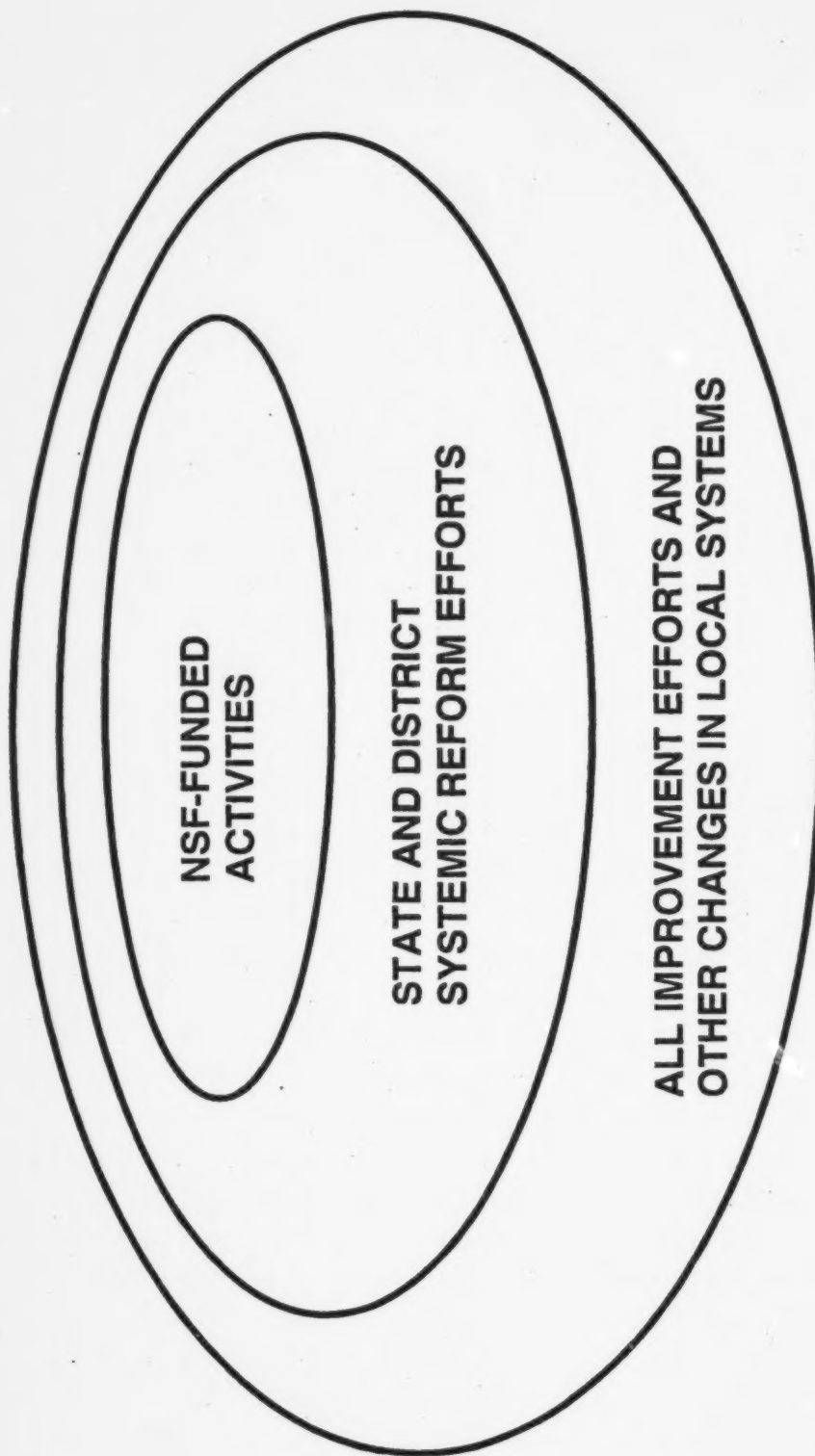
***THE SYSTEM ITSELF IS COMPOSED OF MULTIPLE, HIERARCHICAL, INTERACTIVE AND DYNAMIC SYSTEMS.**

***GOOD INSTRUCTION COMES FROM THE CONVERGENCE OF MANY NECESSARY (BUT NOT SUFFICIENT) SYSTEM SUPPORTS AND EXPECTATIONS.**

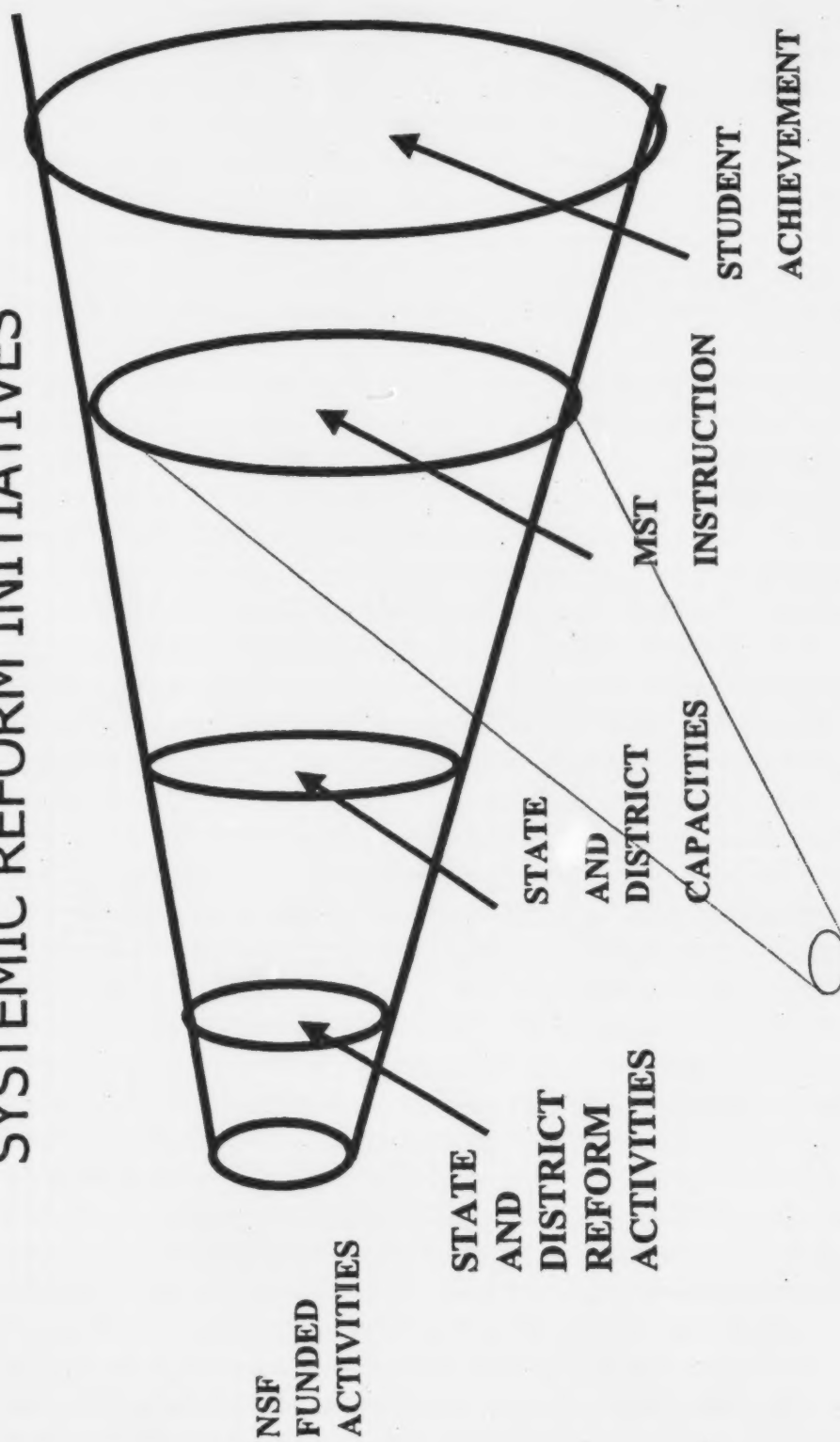
•THUS, THE SYSTEM COMPONENTS MUST ALL BE OF HIGH QUALITY AND "ALIGNED". . . AND

•THE SYSTEM MUST BE PRIMARILY BE IMPROVED BY THOSE LIVING AND WORKING WITHIN THE SYSTEM.

THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES



THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES



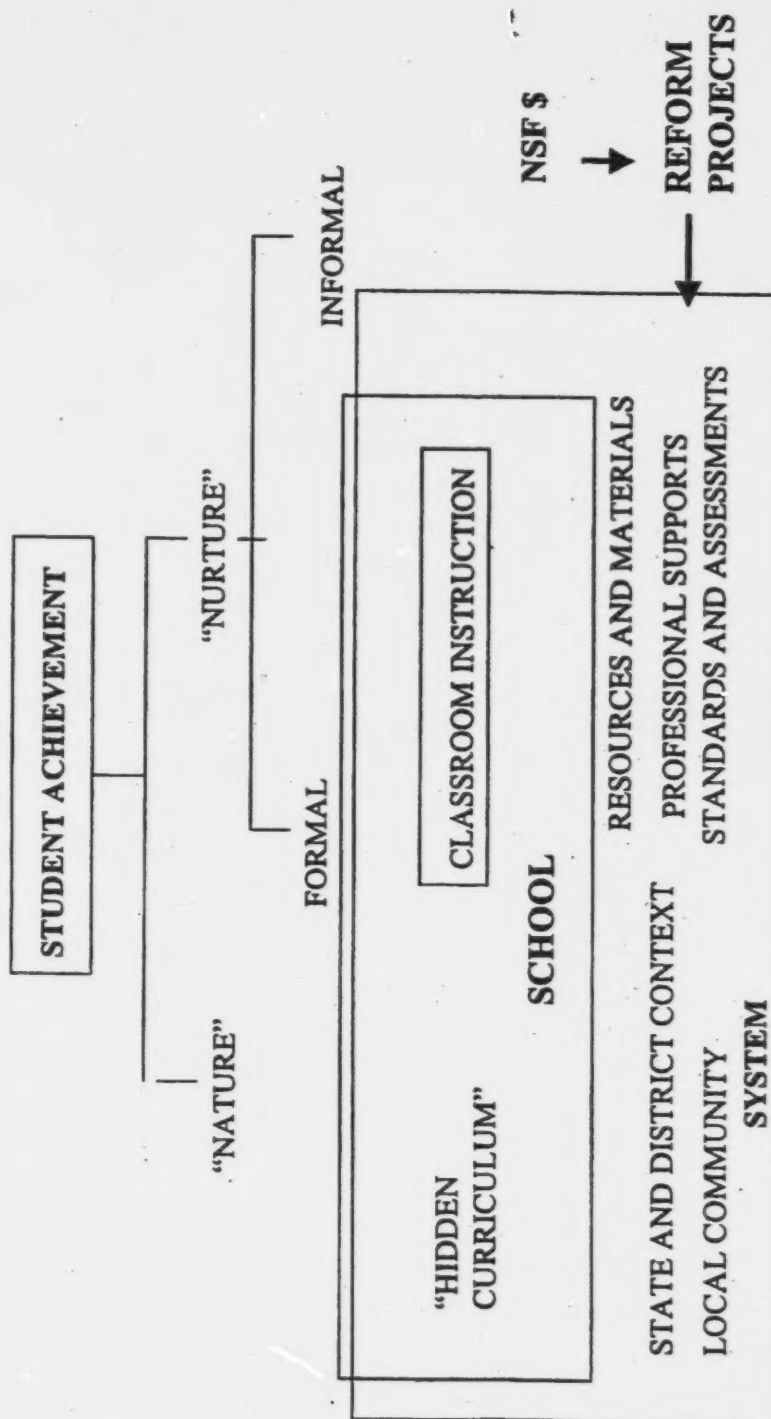
THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES

ISSUES AND LIMITATIONS IN ASSESSING DISTAL IMPACTS

- THE SCALE OF THE INVESTMENT
- THE TIME SCALE OF THE INVESTMENT
- MANY NECESSARY BUT NOT SUFFICIENT COMPONENTS TO ACHIEVE IMPROVED STUDENT ACHIEVEMENT
- THE DYNAMIC AND HIGHLY INTERACTIVE NATURE OF THE SYSTEM THAT IS BEING INFLUENCED
- MULTIPLE SOURCES OF REFORM AND MULTIPLE PROJECTS

==> VERY LOW SIGNAL TO NOISE RATIO

THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES



THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES

$$P(\text{SUCCESS}) = \frac{(L + D + \text{PRI} + \$) \times (PD)}{S + P + F + T + RO + CP} = \frac{(\text{CAPACITY}) \times (\text{DEMAND})}{(\text{BARRIERS})}$$

WHERE

L = Leadership

D = Design, plans

PRI = Policy/Reform Infrastructure

\$ = Discretionary dollars

PD = Political & Public Demand

S = Scale of the System

P = Political change, backlash

F = Financial hardship

T = Turbulence, Instability

RO = Reform Overload

CP = Competing Priorities

THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES

A MODEST PROPOSAL

The appropriate measure of the value of NSF's investments in systemic reform consists of assessing the degree to which the investments have increased the enduring capacity of states, districts and schools to design, initiate and sustain high quality improvement efforts in mathematics, science and technology education.

(END OF STORY...)

THE EVALUATION OF NSF'S INVESTMENTS IN SYSTEMIC REFORM INITIATIVES

NSF is well positioned to fund reform efforts that increase state and district capacity by:

- Developing Leadership and Vision
- Increasing Sophistication of Design (curriculum, professional development, standards...)
- Increasing MST Policy and Reform Infrastructure
- Improving local ability to create public and political demand ?

Introduction to Breakout Session II Question Summary

Each panel was followed by a Breakout Session. Participants were assigned to small groups of ten to twelve, led by a facilitator, in a discussion of three questions and other issues raised by the presenters. Each set of three questions was developed by the organizers of the Forum. At the beginning of the Breakout Session, participants were asked to write their responses to each of the three questions on index cards. The comments that the participants wrote were used to begin the small group discussions. The index cards were given to two people, who provided a synthesis of the conference; comments on the index cards were incorporated into their comments. Responses to the first question are summarized here to provide examples of participants' comments.

Breakout Session II: Successful Strategies for Evaluating Systemic Reform

Participants' Comments:

Q: What are the essential elements of an evaluation of systemic reform?

The 124 participants who responded to this question identified a wide range of approaches for evaluating systemic reform. The grouping of responses given below are listed in order, from issues raised most frequently by participants to those raised less frequently.

1. *Attend to the specific components and indicators of the system and its reform—including forces for change, goals, visions, policy, and accountability.*

Twenty-eight percent of the participants who responded emphasized the importance of measuring and tracking specific components of the system. For example, one respondent wrote, "Need to determine impact on curriculum, professional development, assessment, resources, policies, and student outcomes." Another respondent wrote, "Define/determine all system components. Assess gains (or lack thereof) in all components of system."

2. *Measure student learning and outcomes.* One-fourth of those who responded to the Session II question noted that the evaluator needs to assess student learning. Some of these respondents indicated that an evaluation of systemic reform also should help to identify what constitutes effective measures of student success and achievement. As one respondent wrote, "Student achievement gains mean nothing if the measurement isn't measuring what you really want students to learn. . . . We need to study that part of the system." Some of the responses specifically stated that the focus on student achievement was in contrast to

St. John's perspective. In his talk, St. John stated that reform of systems is so complex and has to be sustained over such a long time that to detect changes in student achievement after only two or three years is premature. At best, evaluations can address improvement in the system's capacity, the degree to which efforts are sustainable, and the issue of whether the system is moving on a trajectory toward significant change.

3. *Evaluate the comprehensiveness and coherence of the system and its reform, as well as the interconnections among the components.* Eighteen percent of the respondents indicated that the evaluation should consider the system as a whole, in contrast to responses that focused on evaluating specific components (see Response 1. above). Some of the respondents indicated that essential to systemic reform is coherence within the system. For example, one respondent wrote, "There must be coherence in the form of the vision and personal understanding to lead the evolution [of systemic change]. . . ." Respondents raised the importance of studying the interactions and relationship among system components as part of evaluating the system as a whole. One respondent replied, "Look at the whole system and system interactions rather than individual components. . . . [Examine] process variables that might relate to standards, curriculum, instruction, . . ." Another respondent felt that the evaluation of systemic reform "must address multiple levels, while acknowledging the transactional and synergistic effects between the components."

4. *Evaluate the processes, means, and conditions needed to attain systemic reform.* Seventeen percent of the respondents indicated that it was important for an evaluation of systemic reform to consider and measure those key attributes of a system that are related to systemic reform, such as improvement of

capacity and alignment. Some of these respondents phrased their response by looking at pressure points considered important for changing the system. One participant agreed with St. John's comments, "Capacity building/ infrastructure development is much more consistent with the immediate scope of the [systemic initiatives] than student achievement. To grab at change in student achievement as artifacts of SI is dishonest at best." Another respondent felt that the main goal for evaluation should be "the extent to which a district has been helped to improve/sustain its own 'reform' efforts."

5. Study beliefs, roles, philosophies, and buy-in of key actors and stakeholders.

Fourteen percent of the respondents made some comment stressing the need to look at the leaders, stakeholders, teaches, and other key actors. For evaluators to do evaluation of systemic reform, they need to understand what these people believe about reform and the vision for reform and what ownership they have in the reform process. It is also important for the evaluator to be sensitive to the fact that stakeholders play multiple roles, or serve multiple constituencies. One respondent noted the importance of determining whether there is a broad base of support among the key stakeholders. Another respondent said it was important for an evaluation to differentiate between what stakeholders say and what they believe. A third respondent thought that an evaluation had to consider the belief system of teachers and how their beliefs relate to their practices.

6. Focus on change over time and the critical indicators that mark change.

Fourteen percent of the respondents indicated that an evaluation should consider progress over time. In order to

do this, it was essential to have baseline information and evidence of how efforts have been sustained over time. One respondent commented that it was essential to collect good base-line data about nearly all aspects of the system and that the evaluation should be designed to monitor the progression of key elements. Another respondent reported as an essential element, "Measuring the change in influence of reform on the school system in the area of policy, curriculum, realignment, standards-based instruction, ..."

Respondents offered other comments on what they regarded as the essential elements of the evaluation of systemic reform. However, no more than 10% of the respondents agreed on any one essential element for evaluation of systemic reform in these remaining comments. Some of the respondents emphasized that an evaluation should be based on a model, or driven by a theory of systemic reform, perhaps using mapping to locate important functions within the system. Some respondents advised evaluators of systemic reform to look at system outcomes other than student learning, such as those related to professional development and change in teaching practices. A few comments cautioned those doing evaluation of systemic reform to consider the time frame for changing school systems and to think about what it is feasible and reasonable to do within a given time frame. Three or four participants mentioned the important role that evaluation can serve by providing feedback to the system undergoing reform and the need for evaluators to consider the different audiences for the evaluation in determining what evidence is gathered and reported. Finally, two participants offered a reminder of how important it is that different evaluators verify findings and valid measures be used.

Panel III: Findings on Systemic Reform from Evaluation and Research

Panel Papers and Authors:

Discovering from Discovery; The Evaluation of Ohio's Systemic Initiative

Jane Butler Kahle

Evaluative Findings on Systemic Reform: Lessons Learned from NSF

Daryl E. Chubin

Value-Added Indicators

Robert H. Meyer

Quantitative and Qualitative Data in the Theory of Systemic Reform

William H. Clune

Discussion Summary and Commentary: Findings on Systemic Reform from Evaluation and Research

Nor-man L. Webb

Panel III speakers presented examples of effective systemic reform efforts and described specific evaluation and research strategies or practices utilized during the evolution of successful systemic reform programs. Presenters' perspectives were framed on one hand by the analysis of one state's experience in evaluating its SSI and on the other by the findings of the National Science Foundation during a decade of supporting systemic change in K-12 mathematics and science. Two presenters focussed on data management and analysis issues that have proved productive in evaluating SSIs.

Jane Butler Kahle, professor of science evaluation at Miami University-Ohio, has been involved both in "doing reform and assessing" it. Former director of Ohio's SSI, she became the principal investigator for *Project Discovery*, which was originally designed to impact middle school science and mathematics education, with a primary focus on the professional development of teachers. Support of *Discovery* was provided from both the state and federal levels, enabling the state to assess changes over a period of five years.

Structurally, a three-tier nested research design was used which yielded different, yet important data at each of three levels:

questionnaires to a random selection of teachers at the state level; annual visits to randomly-selected schools at the district level to validate questionnaire responses; and intensive case studies in five schools (three urban, one small town, and one suburban) focused on equity and school readiness for reform. Two sets of findings of the *Discovery* Project proved to have major policy implications. First, Ohio found compelling evidence that its sustained professional development program significantly changed teaching practices. The average *Discovery* participant reported increased use of standards-based teaching; follow-up questionnaires indicated that these changes were sustained for several years, and that they affected the culture of professional development throughout the state. Second, improved mathematics and science scores were achieved by students taught by teachers who had completed *Discovery's* professional development program; African Americans and white students scored higher than their peers in non-*Discovery* classes. A third major finding in Ohio's project was the importance of effective communication of findings by publishing and widely distributing easy-to-understand charts of results. While mainly quantitative findings were presented, an effort was made to present data in formats the public and the legislature could assimilate.

Kahle made three major points regarding the evaluation of the Ohio systemic initiative: First, it is important to evaluate the reform

while it is occurring, creating a data-driven reform. Ohio found that on-going, continuous evaluation of Discovery attracted the support of legislators and of the parents, teachers, and administrators needed to sustain the effort. Second, evaluation of complex systems must include both quantitative and qualitative data. Third, evaluations intended to guide or accelerate reform are only as effective as their means of communication.

Daryl Chubin, National Science Foundation (NSF) provided a sponsor's perspective on the Foundation's efforts to stimulate systemic reform in mathematics and science in American public education. His paper focused on a rigorous examination of three issues as their importance became defined during NSF's on-going experience with systemic reform: program evaluation as an accountability tool; the measurement challenge; and, quality of information. In confronting the challenge of developing accountability tools, NSF in 1996 produced its *Instrument for Annual Report of Progress in Systemic Reform*, which identified "drivers" that codify the principal dimensions of planning, activities, and reflection in systemic initiatives. The measurement challenge reflects the complexity of the systems with which NSF is partnered in its systemic initiative programs. Because effective, system-wide reform is a complex, nuanced, and uncertain endeavor, with variables that are often specific to the system in transition, no "detached" external measure can produce the insights that careful self-reporting yields. The quality of NSF investment and its payoff in systemic change are a function of the quality of information obtained from SSI projects. A combination of methods to gauge progress includes annual reviews, site visits, and performance effectiveness reviews. Special events such as field hearings also produce vital feedback, commentary, and guidance.

Robert Meyer, research scientist at the Wisconsin Center for Education Research and the Harris Graduate School of Public Policy at the University of Chicago, discussed the weaknesses of the most commonly used educational indicators and the advantages of

employing value-added indicators in the analysis of educational change. After connecting specific measurement flaws and defects to typical indicators, he stated his belief that to have indicators that are appropriate for accountability and/or evaluation purposes, systems need to design indicator/evaluator systems based on the value-added approach. The key challenge is to isolate the contribution of schools to growth in student achievement from all other sources of student achievement over a given time period. Following a critique of the "average test score" as a valid measurement of change, he presented a theoretical simulation. He then proceeded to an analysis of NAEP's 1973-1986 study of 11" grade mathematics scores, showing gains in academic achievement based on average test scores-whereas an analysis of the data based on a gain indicator similar to but not the same as a value-added indicator suggests the opposite. A basic problem is that NAEP data do not permit value-added analysis, since the same students are not sampled for two consecutive NAEP surveys.

Thus, a major question is whether value-added indicators can be used as the foundation for school district, state, and national performance indicator/accountability systems. There are reasons for optimism because researchers/evaluators have been applying value-added models in education and training programs for three decades; and some districts and states have successfully implemented value-added indicator systems. The value-added approach to measuring school performance relies on a statistical model to identify the distinct contributions made by schools to growth in student achievement. In conclusion, he indicated that four factors determine the quality of value-added indicators: testing frequency; the quality and appropriateness of the tests; adequacy of control variables included in the statistical models; and, the technical validity of the statistical models used to create the indicators. To implement a value-added system, states and schools need to consider testing students at every grade level, including summer school and in-migrating students; it is important that states make it a major priority to collect

extensive, reliable data on student/family characteristics and that they develop tests that are technically sound and aligned with educational goals.

William H. Clune, professor of law at the University of Wisconsin and project co-director at the Wisconsin Center for Education Research, **focussed** on one major methodological aspect of using quantitative and qualitative data in evaluating systemic reform. Using a theory developed by a NISE team, he demonstrated the usefulness and limits of quantitative ratings as a tool for understanding systemic reform. Based on nine SSI case studies published by SRI in March, 1998, he rated the states on the basis of four variables—each measured by breadth and depth—that are essential aspects of systemic change: reform, policy, curriculum, and achievement. Attributes of successful reform included vision, strategic planning, networking with policy makers and with professionals, institutionalization of the reform structure, leveraging of resources, and public outreach and visibility. The reform was considered broad if it included all of these elements and the elements touched all levels of policy, and deep to the extent that each element was well developed and influential. He explained that to test the theory of systemic reform, it was necessary to determine whether higher levels of reform and policy do, in fact, produce change in teaching and learning. Low ratings in reform and policy indicated, however, that as a group the SSIs fell short of achieving the ultimate goal of transforming entire states.

Clune developed several points regarding the usefulness of quantitative data in making

generalizations about reform, pointing out, however, that qualitative information is needed as a means of interpreting the numbers. A common substantive problem was the pedagogical orientation of reform—its emphasis not simply on teaching, but on active learning. Direct means of influencing curriculum were relatively rare, especially early in the reform process. The gap between pedagogy and content narrowed as reform progressed, partly, he noted, as a result of prodding by NSF. But few of the SSIs began with changes in course content and pedagogy embedded in their development design. The best evidence of curriculum change has come from data on teacher training, surveys of teacher attitude and practice, and some cases of whole-school restructuring. Early in the reform effort, classroom change was not a clear objective of policy; many SSIs were built around professional development, with teacher capacity as the goal, rather than curriculum upgrade.

Briefly discussing the limits of usefulness for numbers, he noted that many generalizations about systemic reform require purely qualitative analysis. He concluded with the statement that no other technique seems as capable as numerical ratings of testing the basic hypothesis that strength in one variable produces strength in the next, summarizing the overall progress of individual reforms and the reform effort, analyzing the status of reform components across sites; but that many data in a quantitative analysis are the result of qualitative inquiry and that many important patterns across reforms can only be recognized and understood as a result of thoughtful qualitative inquiry.

DISCOVERING FROM *DISCOVERY*: THE EVALUATION OF OHIO'S SYSTEMIC INITIATIVE

Jane Butler Kahle
Miami University-Ohio

Expand the Discovery Project. This National Science Foundation-funded initiative started in 1991 as a project devoted to improving middle school science and mathematics education. The primary focus of the program has been on teacher professional development. In 1997, this successful program received appropriations of \$2.5 million per year from the state budget and the mission was expanded to improving math and science education from elementary through graduate school. The Taft Administration will expand this program and refocus it on the elementary and middle school years.

(Bob Taft, Ohio Governor-Elect, September 2, 1998, p. 29).

Background

Discovery began in 1991, when Ohio was among the first cohort of states to receive Statewide Systemic Initiative awards from the National Science Foundation (NSF). During the past decade, we have simultaneously been *doing* reform and *assessing* it. Although other papers detail specific findings (see, for example, Kahle, 1997; Boone, 1998; Carnes, 1998; and Damjanovic, 1998), this one will focus on some findings that have had major policy implications.

Discovery has been fortunate in two ways: first, Ohio's General Assembly continued to support *Discovery* after the period of NSF funding; and, second, NSF funded a new project, *Bridging the Gap: Equity in Systemic Reform*, that continues the evaluation of systemic reform in Ohio. Therefore, we have been able to assess changes over five years and to use our findings to guide and accelerate the reform of mathematics and science education in Ohio. Further, we anticipate that the catalysts and barriers identified in Ohio will be common to many systemic efforts and will contribute to the knowledge base about systemic reform.

Three years into Ohio's reform, we began to assess its progress. A three-tier, nested research design has been used, which yields different, yet important, data at each of three levels. At the state level, we have used questionnaires with a random sample of

teachers and administrators in over 100 schools to provide evidence of changes in teaching practice, in administrative support, and in teacher expectations. At the district level, annually we have visited from 12 to 16 schools that are part of the larger random sample. Our observations over several days have validated questionnaire responses and have allowed us to place the quantitative data in context. In addition, student achievement and attitudinal data have been collected in the schools visited. Simultaneously, we have been conducting intensive case studies in five schools (three urban, one small town, and one suburban). The case studies are focused on equity and on how systemic reform works in schools that are at different stages of readiness for reform. They are providing information about opportunities to learn as well as about catalysts and barriers to reform.

Discoveries About the Evaluation of Systemic Reform

For this paper, I have identified three major points gleaned from the evaluation of one systemic initiative. First, it is important to evaluate the reform while it is occurring. In that way, policies and practices may be influenced by the findings, and the reform becomes data-driven. In addition, the ongoing and continuous evaluation of *Discovery* has helped it attract the support that is needed from state legislators and governors and, more

importantly, from parents, teachers, and administrators to sustain a reform. Second, evaluation of complex systems must include both quantitative and qualitative data. And, third, evaluations intended to guide or accelerate the reform process are only as effective as their means of communication. These points are illustrated through the findings presented below.

Discovery's Most Powerful Findings

Although *Discovery* and *Bridging* have assessed multiple aspects of Ohio's reform, two sets of findings have had a major impact on the policies and practices of the reform. First, there is compelling evidence that *Discovery's* sustained professional development (six week, summer, content institutes, taught by inquiry, followed by six academic year seminars on equity, assessment, and pedagogy) have changed teaching practices. Teachers completed questionnaires regarding the nature of their teaching before they began their summer professional development and in the spring of the following year for three years. The items reflected a range of standards-based teaching practices (e.g., working in small groups; doing inquiry activities, making conjectures, and exploring alternative ways to solve a problem). The average participant, in both mathematics and science, reported an increase in the use of standards-based teaching practices after participation in the SST's professional development program, and follow-up questionnaires indicated that those changes were sustained for several years (Supovitz, 1996). These findings have been corroborated by classroom observations, teacher and student interviews, and by teacher and student portfolios. Other evidence indicates that they have affected the culture of professional development in Ohio as more districts offer or reward long-term experiences.

The second set of important findings involves student achievement. Because it is difficult, if not impossible, to establish causality in a complex, multi-year reform effort, we have analyzed student achievement data on *Discovery's* Inquiry Tests in several

ways, and we have sought other types of achievement data. The intent is to identify patterns or trends that suggest that any change is more than a chance phenomenon. The examples below illustrate several strategies used.

In one analysis of student achievement, socio-economic level of the students was controlled; while, in another one, possible bias in the teacher group was controlled. Student achievement data in both studies indicated improved learning by African American and White students who were taught by teachers who had participated in *Discovery's* sustained professional development. For example, a comparison of 610 science students in matched science classes (e.g. seventh grade life science) indicated that both African American girls and boys in classes taught by *Discovery* teachers scored 9% higher on the *Discovery* Inquiry Test in science than did their peers in the matched classes. In addition, White girls in *Discovery* classes scored 10% higher, and White boys scored 4% higher than their peers in *non-Discovery* classes (Damjanovic, 1998). Similar results have been found in mathematics classes (Goodell, 1998).

Other analyses have controlled for teacher motivation, or the "volunteer" effect, by comparing the achievement of students whose teachers have volunteered to participate, but have not done so, with that of students whose teachers have completed the professional development. (See Corcoran, Shields, & Zucker, 1998, for a discussion of this issue). The positive effect of the *Discovery* professional development is suggested by higher scores (from 2% to 7%) on both the mathematics and science Inquiry Tests by students (N = 2374) whose teachers had completed the *Discovery* programs, compared to those who had not (Supovitz, 1996).

Independent analyses have established that the gender gaps in both mathematics and science have been decreased both across and within racial groups (Damjanovic, 1998; Goodell, 1998). Analyses, using the whole data set, indicate that the achievement gap between African American and White students (favoring Whites) has narrowed but

persists. However, those analyses do not necessarily involve African American and White students who are in the same classroom. Therefore, we analyzed the achievement of African American and White students in the same classes by using only classes in the sample that had at least 25% of their students in a minority group (either 25% African American or 25% White students). Although many classes did not fit that profile, we had a representative sample (comparable numbers of classes taught by *Discovery* and *non-Discovery* teachers) for three years in mathematics. One hundred and eight classes were involved, enrolling over 3000 students. The findings show a narrowing of the achievement gap (favoring Whites) in mathematics in classes taught by *Discovery* teachers (from 10.4 percentage points in 1995 to 7.5 in 1997), and a widening of the gap (from 7.3 percentage points in 1995 to 15.1 in 1997) in classes whose teachers had not participated in the sustained professional development.

Two new types of achievement data have been obtained. First, in 1998, we were able to obtain Ohio Proficiency Test (OPT) mean scores for 1997 and 1998 for schools in several large urban districts. Because those scores are reported publicly by pass/fail rates only, they have not been useful in the past. Two criteria were used to select the middle/junior high schools in each district for the analysis. They were: over 70% African American students and over 55% students eligible for free or reduced-price lunch. All schools in an urban district that met those criteria were included in the analysis, and each district was analyzed separately. Analyses were run for whole districts in order to explore any effect of a "critical mass" of *Discovery* teachers on OPT scores in mathematics and science. "Critical mass" was operationally defined as over 5 1% (full time equivalent) of the science and mathematics teachers in the school had participated in a *Discovery* professional development program. Over 13,000 seventh through ninth grade students attended the schools that were used in the analysis. We identified two patterns. In districts where policies were aligned with

reform practices, OPT scores in mathematics and science rose with the percent of *Discovery* teachers, while in districts with little policy alignment the percent of *Discovery* teachers did not affect OPT scores. For example, in high alignment districts, OPT scores improved 17.5% in mathematics and 9.2% in science in schools with more than 5 1% *Discovery* teachers. On the other hand, in those same districts OPT scores declined 11.3% in mathematics and 3.3% in science in schools with fewer than 25% *Discovery* teachers. In schools with little alignment among state, district and *Discovery* efforts, there was little variation among the scores of students by percentage of teachers who had participated in the professional development. We were able to interpret and place these findings in context because of the extensive amount of qualitative data we had collected in the cooperating districts.

Not all of our attempts to obtain evidence concerning achievement have been successful. In 1998, we explored the use of performance assessments by implementing performance tasks from the Third International Mathematics and Science Study (TIMSS) in selected schools (student N = 500). In addition, multiple choice versions of selected TIMSS' tasks were added to the *Discovery* Inquiry Test. For one school, student responses (N = 65) on the TIMSS multiple choice items were compared to their responses on the TIMSS performance tasks. Initial analysis of the data suggests that paper and pencil tasks alone inadequately measure student understanding, particularly the understanding of urban, African American students (Kelly & Kahle, 1998). For example, for those students who responded to both types of items (multiple choice and performance), 86% were able to identify patterns in data when they had collected the data and drawn the graph (performance task). Only 8% were able to correctly identify patterns when the data were presented in the paper and pencil test. However, expense as well as unresolved technical problems in both delivery and scoring have prohibited the continued use of performance items.

My third major point was the importance of effective communication. Annually, our findings are published and widely disseminated in the *Pocket Panorama*. To date we have communicated mainly quantitative findings, learning how to present data in easy to understand ways and how to reach key legislators. The next challenge is to learn to communicate succinctly and clearly the complex stories that are emerging from our case studies.

Summary

Analyzing data in multiple ways has allowed us to tell a convincing story—a story that has led to substantive changes. First, the culture of professional development has changed in Ohio, with long-term, substantive programs preferred or mandated. Second, in order to accelerate improved student achievement, we have moved from teachers to schools as the unit of change. Variations of the content institutes are taught now at the district level, and Discovery's new institute for principals is in demand across the state. Evaluations that occur concurrent with reform, that collect multiple types of data, and that effectively communicate their findings with broad audiences can shape a reform.

The preparation of this paper was funded in part by a grant from the National Science Foundation, Grant #REC 9602137 (J. B. Kahle, Principal Investigator) and by National Science Foundation Grant #OSR-92500 (J. B. Kahle and K. G. Wilson, Co-Principal Investigators). The opinions expressed are those of the authors and do not necessarily reflect the position of the National Science Foundation.

References

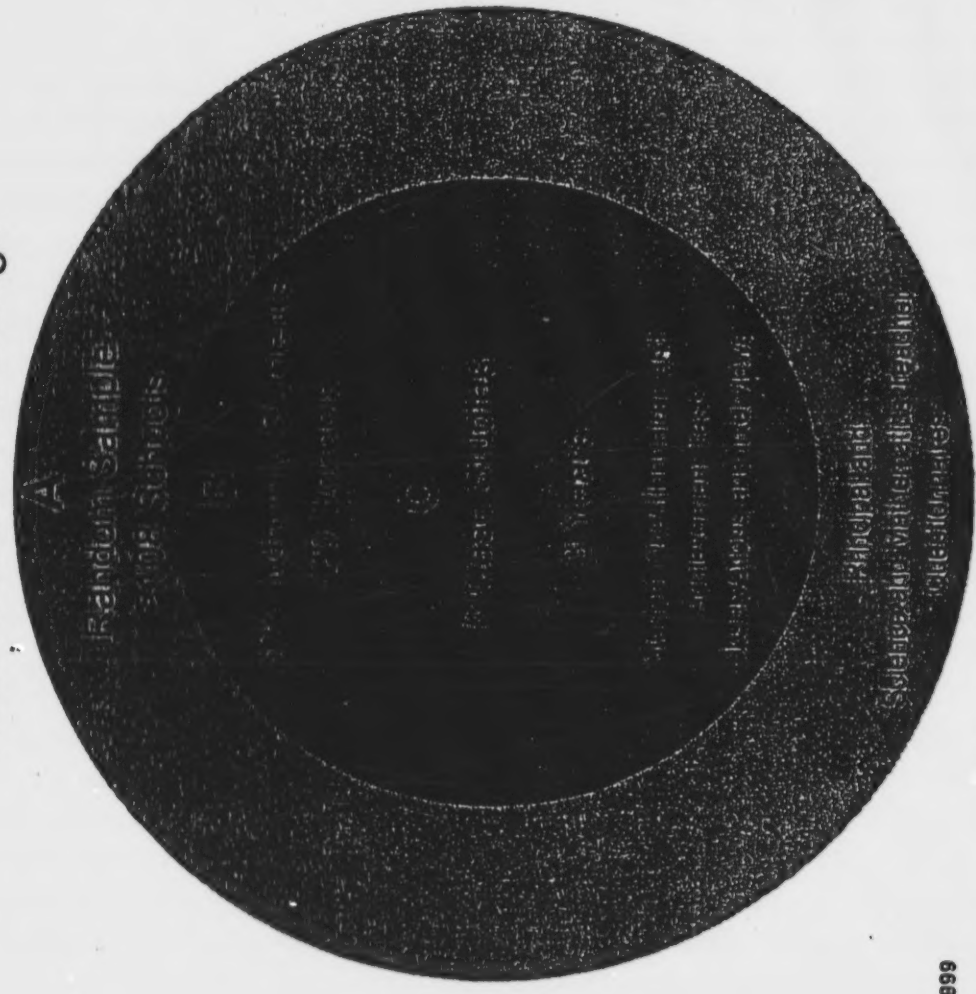
- Boone, W. J. (1998). Assumptions, cautions, and solutions in the use of omitted test

- data to evaluate the achievement of underrepresented groups in **science**—Implications for long-term evaluation. *Journal of Women and Minorities in Science and Engineering*, 4, 183-194.
- Carnes, G. N. (1998). An assessment of equitable instruction within three urban middle school classrooms. *Journal of Women and Minorities in Science and Engineering*, 4, 283-296.
- Corcoran T. B., Shields, P. M., & Zucker, A. A. (1998). *The SSIs and professional development for teachers*. Menlo Park, CA: SRI International.
- Damnjanovic, A. (1998). Ohio Statewide Systemic Initiative (SSI) factors associated with urban middle school science achievement: Differences by student sex and race. *Journal of Women and Minorities in Science and Engineering*, 4, 217-233.
- Goodell, J. E. (1998). *Equity and reform in mathematics education*. Unpublished doctoral thesis, Curtin University of Technology, Perth, Western Australia.
- Kahle, J. B. (1997). Systemic reform: Challenges and changes. *Science Educator*, 6, 1-6.
- Kelly, M. K., & Kahle, J. B. (1998). *A comparison of student achievement on performance and paper-and-pencil assessment tasks*. Unpublished manuscript, Miami University, Oxford, OH.
- Supovitz, J. (1996, December). *The impact over time of Project Discovery on teachers' attitudes, preparation, and teaching practice. Final report*. Chapel Hill, NC: Horizon Research, Inc.
- Taft, R. (1998, September 2). *The Taft high-tech education strategy: Improving math and science performance*. Unpublished manuscript, Office of the Governor-elect, Columbus, OH.

Discovering from *Discovery*: Three Major Points

- ◆ Evaluation must be ongoing and occur during the reform process, so that the reform becomes data driven.
- ◆ Evaluation of complex systems must include both quantitative and qualitative data.
- ◆ Evaluations intended to guide or accelerate the reform process are only as effective as their means of communication.

Figure 1
Bridging the Gap: Equity in Systemic Reform
Nested Research Design



© Jane Butler Kahle, 1999

Bridging the Gap: Equity in Systemic Reform

Student Achievement Results for Students in Matched
Classes Within the Same School (1995)

	Science Achievement by Classroom Type		Difference
	SSI Teachers	Non-SSI Teachers	
African American Females	43.9%	35.0%	8.9%
Males	40.0	30.6	9.4
White Females	59.5	49.7	9.8
Males	54.3	49.6	4.7

N = 610

Source: Damjanovic, 1998

© 1999, Jane Butler Kahle

Bridging the Gap: Equity in Systemic Reform

Predicted Student Performance on *Discovery's*
Inquiry Tests In Mathematics and Science*

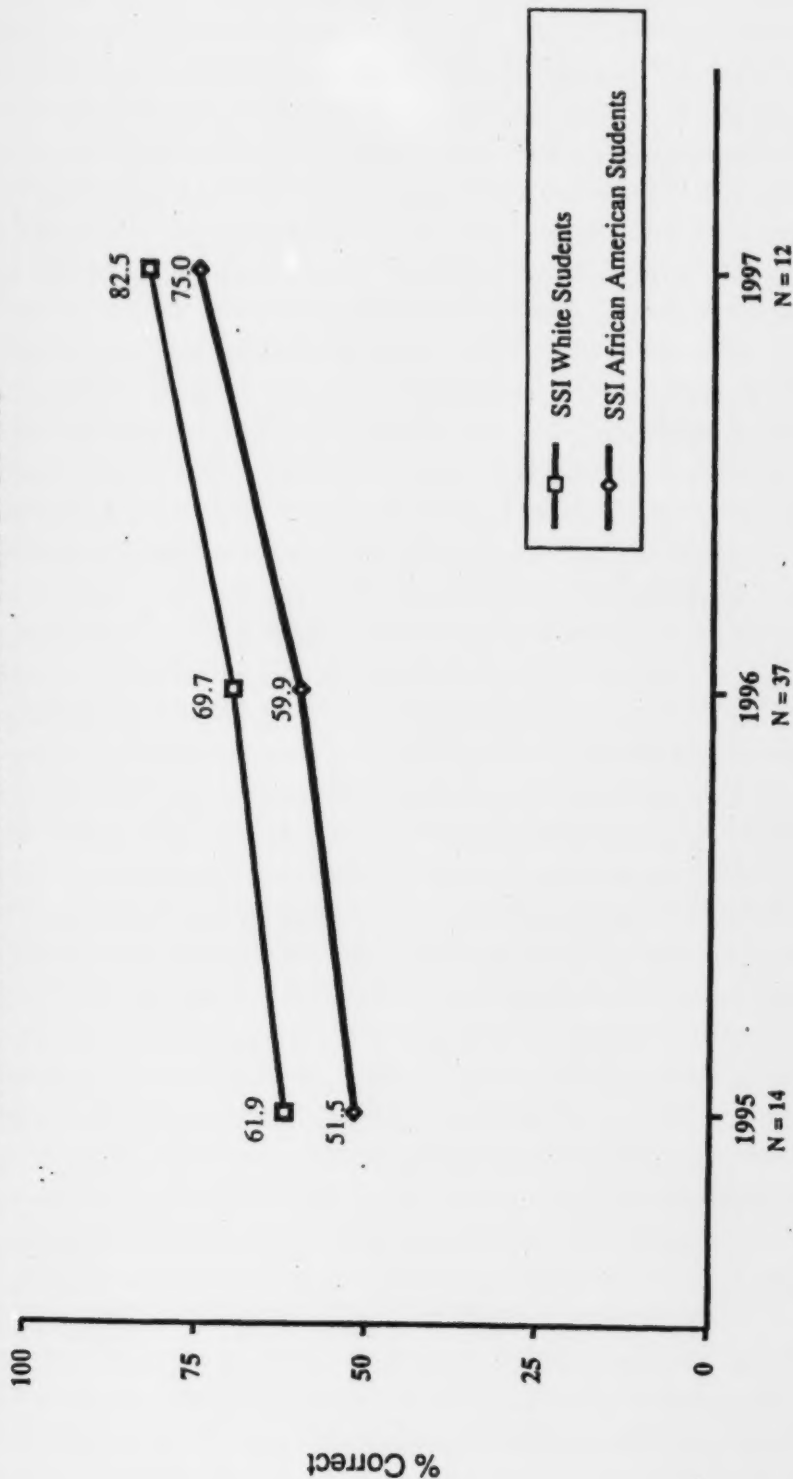
			Student Achievement by Classroom Type		Difference
			SSI Teachers	SSI Applicant Teachers	
Science (n=1,144)					
African American	Females	46.9%	43.7%	3.2%	
	Males	44.2	41.0	3.2	
White	Females	62.3	59.1	3.2	
	Males	59.6	56.4	3.2	
Mathematics (n=1,230)					
African American	Females	54.2	47.3	6.9	
	Males	54.3	52.3	2.0	
White	Females	66.2	59.2	7.0	
	Males	66.1	64.2	1.9	

*Controlled for urbanicity, concentration of poverty and grade level.

Source: Supovitz, 1996, February

© 1999, Jane Butler Kahle

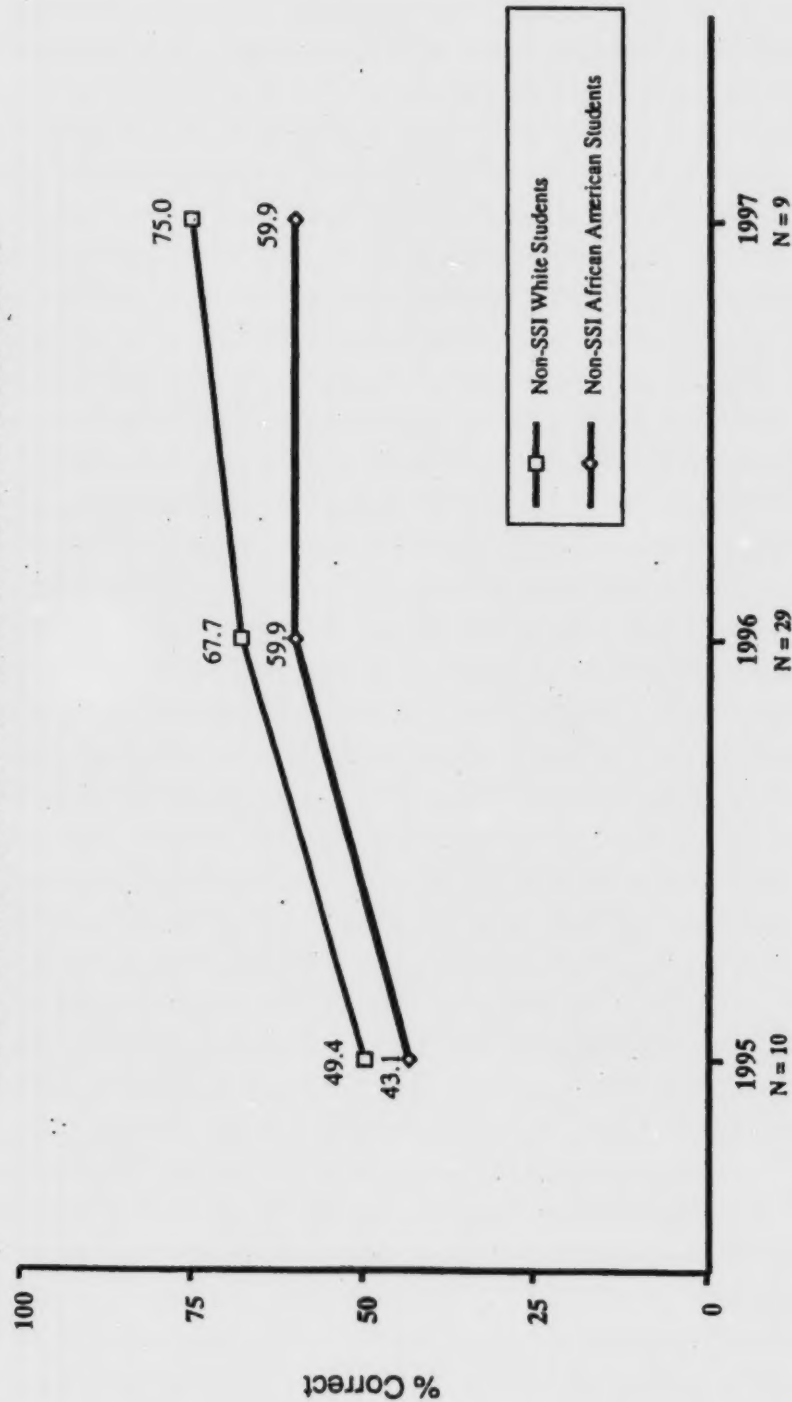
Bridging the Gap: Equity in Systemic Reform Achievement in Mathematics for African American And White Students in SSI Classrooms



Note: Achievement measured by *Discovery Inquiry Test* in mathematics.

Bridging the Gap: Equity in Systemic Reform

Achievement in Mathematics for African American
And White Students in Non-SSI Classrooms

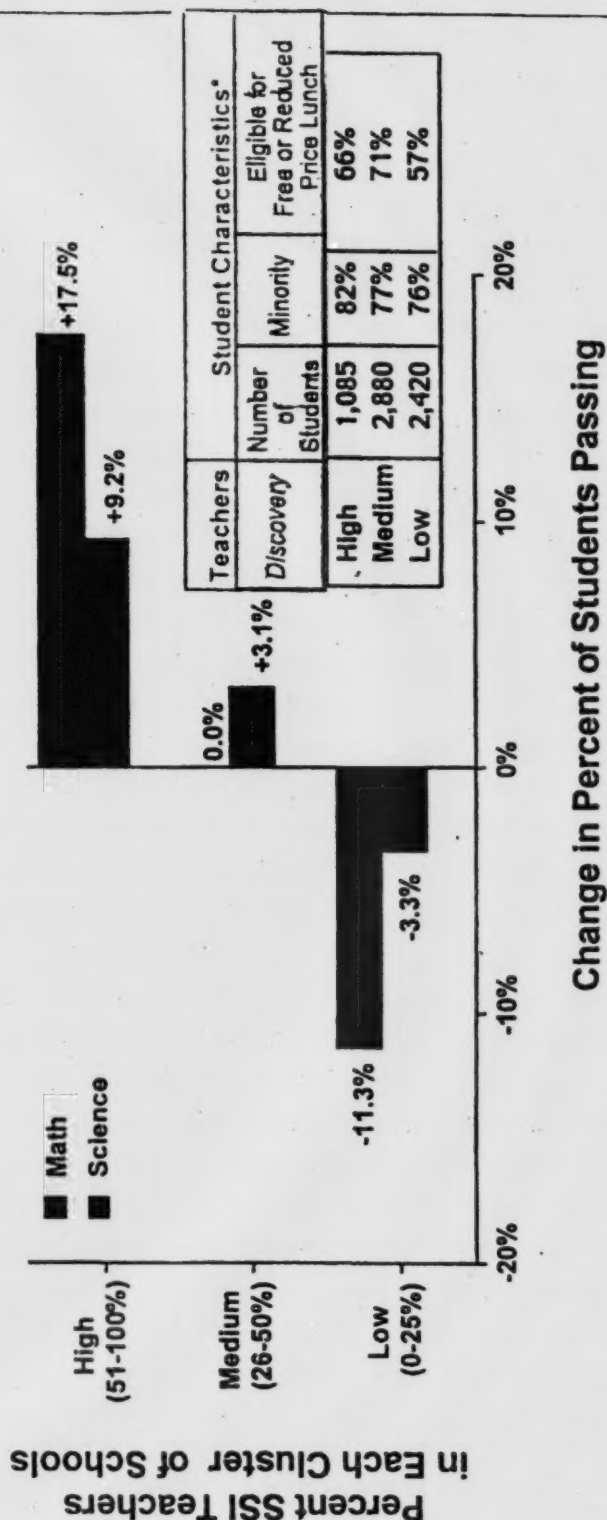


Note: Achievement measured by *Discovery Inquiry Test* in mathematics.

© 1999, Jane Butler Kahle

Bridging the Gap: Equity in Systemic Reform

Changes in Percent of Students Passing Ohio Proficiency Tests
(Grade 8) in Cities with Aligned Reform Policies

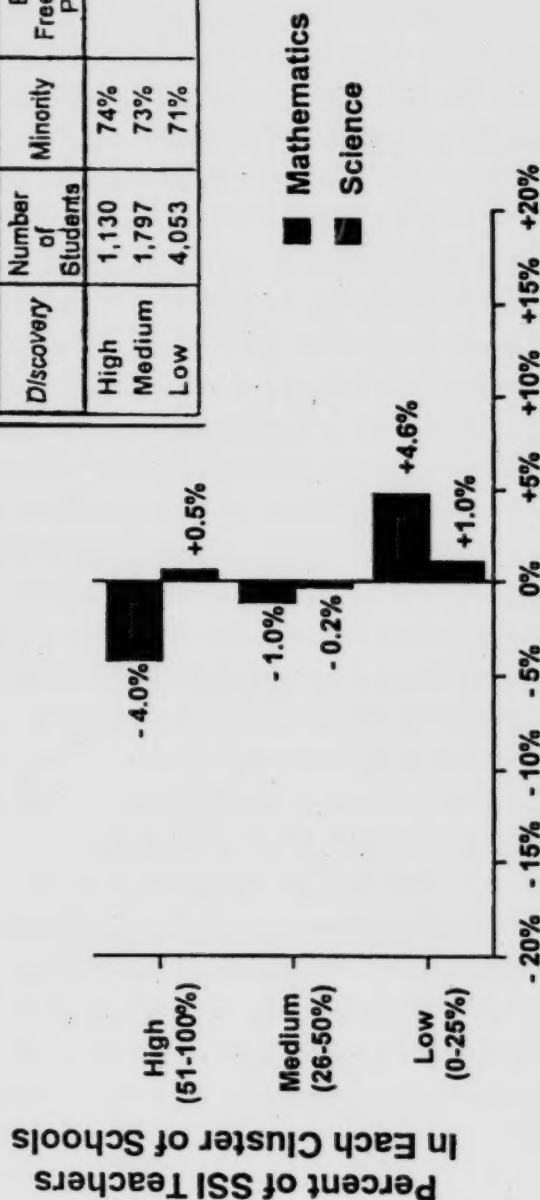


*The average percent for all schools in each cluster is based on Ohio Department of Education data.

© 1999, Jane Butler Kahle

Bridging the Gap: Equity in Systemic Reform

Changes in Percent of Students Passing Ohio Proficiency Tests (Grade 8) in Cities with Varied Reform Policies



Teachers Discovery	Student Characteristics*		
	Number of Students	Minority	Eligible for Free or Reduced Price Lunch
High	1,130	74%	57%
Medium	1,797	73%	63%
Low	4,053	71%	76%

Changes in Percent of Students Passing (1997-1998)

*The average percent for all schools in each cluster is based on Ohio Department of Education data.

EVALUATIVE FINDINGS ON SYSTEMIC REFORM: LESSONS LEARNED FROM NSF

Daryl E. Chubin
National Science Foundation

A Sponsor's Perspective

The Statewide Systemic Initiative (SSI) was the National Science Foundation's (NSF) inaugural effort to stimulate systemic reform. It was an experiment, an alternative to education-as-usual, and in keeping with the culture of the Foundation, a merit-based competitive program featuring formidable pre- and post-award performance requirements.¹ Our assigned role in this panel is "to present a few overarching findings from the Systemic Initiatives [SIs]." While we will do that, what may be more valuable is to indicate how we arrived at those findings and how we as a sponsoring agency use various data to refine our understanding of ongoing change in those educational sites.

The Foundation has itself learned much since 1991 and improved as a partner, a technical assistant, and a codifier of what it takes to move an entire system toward reform of its mathematics and science instruction. We are not alone. Reflecting on our stewardship, as one of us did in testimony before the House Committee on Science last summer,² several impressions became clear. Three overarching factors frame our remarks:

1. NSF has not told the story of its pre- and post-award procedures especially well.

¹ An NSF perspective on this experiment was first offered in D. E. Chubin, "Systemic Evaluation and Evidence of Education Reform," in D. M. Bartels and J. O. Sandler, eds., *This Year in School Science 1994—Implementing Science Education Reform: Are We Making an Impact?* 1997, pp. 121-142. Washington, DC: American Association for the Advancement of Science, 1997.

² D. E. Chubin, Testimony before the Subcommittee on Basic Research, the Committee on Science, U. S. House of Representatives, July 23, 1998.

2. Systemic reform means different things to different people, not all of whom are convinced that the unit of analysis—the system—is amenable to wholesale change.
3. Evaluation data have many origins and can be challenged on various grounds.

Suffice it to say here that recipients of NSF support—the reform sites, the SI evaluators, and researchers who glimpse all of this from some distance—are not seeking the dollars alone, but something else: the integrity of the pre- and post-award process, the imprimatur of NSF's quality control, and a support network of participating Systemic Initiative awardees. The Foundation, for example, has helped states create an infrastructure—the conditions for continuous improvement—but we alone cannot spur teachers' willingness to augment their knowledge and skills, or change their classroom practice.

Accountability Tools

As an accountability tool, program evaluation serves many purposes that are typically distinguished as "formative," or periodic monitoring, and "summative," or ex post facto judgments about the program as an investment strategy. NSF has been committed to both, under a 1992 Congressional mandate. This occasioned a third-party evaluation of the SSI program,³ but it certainly did not preclude various site-based project

³ SRI International, A Report on the Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program, prepared under contract, June 1998 (NSF 98-147).

evaluations and agency-centered data-collections and analyses. Indeed, one of the most important developments in 1996 was the establishment of NSF's *Instrument for Annual Report of Progress in Systemic Reform*. This instrument identified six critical program areas that drive systemic reform, referred to as "drivers," which not only highlight the major elements for all site-generated reports and agency reviews, but also codify the principal dimensions for all planning, activities, and reflection by any systemic initiative funded by NSF. Since 1966, these six drivers have served as the formal annual accountability organizer for the SIs.

Some might also say that NSF's SSI Phaseout and Phase II processes are the clearest statement of what the agency has learned in the seven years of the SSI program. Put another way, NSF's vigilance precludes any whimsical or arbitrary decision. Review is ongoing with information flowing both to the states and to NSF from multiple sources.

Taken together, the following key findings have emerged as vital for the successful implementation of systemic reform. The lessons harvested by NSF include:

- Different states have used distinct pathways to achieve higher performance standards in science and mathematics.
- Systemic reform requires sustained and accomplished management of forces, monied interests, and conflicts that impede improving student achievement. This requires systemic thinking.
- Inadequacy and nonexistence of a reform-capable infrastructure inhibits "traction," or incremental progress in reform.
- There is an exceedingly small pool of expertise resident in or external to the awardee systems.
- The baseline in states is minimal, so the trajectory of comprehensive reform is longer than a five-year venture.

- NSF and site-based accountability requirements help to propel reform and functions as an incentive for performance improvement.

NSF would argue that the SSI program has catalyzed reform and leveraged scarce resources in support of standards-based mathematics and science teaching. Our process was hailed at the sites of reform, by our many partners and communities struggling with school change and student performance. Criticisms of the Foundation focus on management inconsistency and unrealistic expectations. What even critics admit, however, is the soundness of the concept, which some states (e.g., Florida) pursued after NSF withdrew. What most will never acknowledge is that an outside interest was needed to jump-start a stalled or nonexistent reform effort. NSF's self-assumed role was to perturb the system as we saw it—and assist the state to effect the changes it proposed. This same approach, and accountability regimen, guided NSF's Urban Systemic Initiative (USI) program.

The Measurement Challenge

That change is uneven across districts and schools should not be surprising. To attribute that limited success to design flaws or insufficient incentives, or on the analytical side, to evaluation methodologies that fail to capture the complexity or efficacy of what was or was not happening, however, is the nub of this panel's challenge: How do evaluations distinguish real change from measurement error? How do they attribute causes to observed effects? Who do they hold accountable for the successes and failures? These are not merely intriguing research questions. They reflect high-stakes political measurements and interpretations that can do harm as well as good, allocate blame as well as credit, and serve to nudge or retard progress in perceiving and promoting system-wide education reform.

Perhaps the biggest testament to NSF's investment in the SSI program has been the

application of lessons learned to the Urban Systemic Initiatives, established in 1993.⁴ Some of these lessons have been collected in handy, accessible booklets that caution both practitioners and "data-handlers" (a.k.a. "researchers" and "evaluators") about the movements among theory, practice, measurement, and interpretation of systemic reform. For example, in a recent issue brief, NSF's systemic initiative technical assistance provider summarizes NSF and site-based learning about the need for disaggregated data: Once data have been disaggregated and differences are found, the next question is What do those differences mean? How does the information provided by disaggregating data help an SI reach its goals?

The process is a little bit like unraveling a ball of yarn. Sometimes the process is simple and goes very quickly; at other times, it is necessary to stop and untie a number of knots along the way.

Suppose, however, that no enrollment differences had been found despite well documented group differences in performance. Should such a finding be interpreted as showing that coursetaking was not a problem? The answer is "yes" and "no." The finding may be a signal that courses are providing different opportunities to learn despite their common labels. Further examination of data may be needed to determine whether the differential performance is related to school characteristics, teacher characteristics, or other factors.⁵

⁴ The USI program was designated as a comprehensive effort to promulgate fundamental change in the quality and level of K-12 mathematics, science, and technology education in 28 U.S. cities having the largest number of school-aged children (ages 5-17) living in poverty, as determined by the 1990 Census. Urban school systems enroll approximately half of all school-age children in the United States, and disparities in academic performance between these students and their counterparts in suburban schools persist.

⁵ National Science Foundation Systemic Initiatives, *A Brief Primer on Disaggregation of Data*, Issue Brief 1, undated, p. 12.

Similarly, NSF's SI sites formulated models of urban K-12 reform that were collected and organized by the technical assistance provider.⁶ Conventional wisdom suggests that such publications are suspect due to their self-report origin, though articles in the mass media and education literatures rely heavily (albeit not exclusively) on such reports,⁷ which are now commonly produced and disseminated to inform as well as promote reforms. Are reports such as *A Pocket Panorama of Ohio's Systemic Reform, 1998* inherently flawed because it has not undergone journal peer review? Is a private communication to NSF, such as the Puerto Rico SSI Program Effectiveness Review Presentation (Dec. 18, 1998), less credible because it was prepared as required by a sponsor? Finally, does a contracted five-year program evaluation report have more or less veracity because of the insider status of its data, its authors, or its interpretations?

These are not questions readily answered. Each must be taken on the merits of its content. One factor must be weighed in any analysis: effective system-wide reform is a complex, nuanced, and uncertain endeavor, with variables that are often highly specific to the system in transition. Accordingly, no "detached" external entity can secure the insights that careful self-reporting yields. Further, the body of knowledge that applies to effective reform is evolving rapidly. Arguably, more relevant expertise about transition processes resides in the practitioner community than in the traditional academic research and/or evaluation community. For both of these reasons, the need for balancing external evaluation with self-reporting is critical and the definition of "balance" itself is complicated.

⁶ Westat*McKenzie Consortium, *The National Science Foundation's Urban Systemic Initiatives (USI) Program: Models of Reformed K-12 and Mathematics Education*, October 1998.

⁷ For an example of a hard-hitting and multisource feature, see J. Mervis, "Mixed Grades for NSF's Bold Reform of Statewide Education," *Science*, 282, Dec. 4, 1998, pp. 1800-1805.

Quality of Information-Quality of Investment and Payoff

To revisit the original issue: How should independent vetting of site-based data proceed? To gauge the progress of projects, NSF employs a combination of triangulating methods, including annual reviews, site visits, and performance effectiveness reviews. Each encompasses different subsets of NSF program and EHR senior staff. All offer information on SI performance from a variety of perspectives. These are synthesized and discussed as an input to the site's renegotiated cooperative agreement. They also inform discussions by EHR senior staff of program performance and funding levels to propose in the coming fiscal year budget.

In addition, special scrutiny comes in the form of events such as the July and October 1998 field hearings held in Chicago, IL, and San Juan, PR, by the National Science Board (NSB), the governing body for NSF. The hearings focused, respectively, on city and state school-based education reform. Witnesses were drawn from research, evaluation, practitioner, and private sector sponsor **communities** to augment NSF's standard data collections and allow NSB members to exchange views with local and national stakeholders.⁸ Soliciting perspectives on investments and their payoffs produces hints for management that are evaluative, but not formally labeled as such. For example, in Chicago, NSB members and NSF staff heard:

Respect the centrality of **content**—teachers must know it to teach it—and the fit between what is taught and what is tested. The choice of curriculum materials may be less important than the teaching culture, which stresses “don't leave anything out,” a finding reinforced by the **TIMSS** results on the lack of focus and coherence.

“Action research” that combines a negotiated research design, an intervention strategy, and feedback to participants is of

special value. But formidable attribution problems (notably school-based v. **out-of-school** influences) exist.

A corporate partner to reforming school systems in **USI** cities lauds the consistency that a **5-year** award helps to bring. Nevertheless, the media and the public must engage the notion that reform takes time. Parents especially must become more critical consumers of reform.

While not derived from formal program evaluations, such insights represent important feedback, commentary, and guidance to stewardship. Do we defend the purity of the labels or utilize the information for the benefit of the particular site and the program as a whole? For NSF, the choice is clear.

Prospects

The upshot of NSF's accountability regimen is that “program evaluation,” as traditionally defined, represents but one tool among many. And while important for documenting via a third-party and therefore asserting an independent view of program performance, it may not be the most timely, accurate, or incisive vehicle for learning about, adjusting, managing, or interpreting what goes on in a Systemic Initiative. No account is definitive and there is no readily identifiable evaluation community steeped in systemic reform. Without a “prevailing community standard,” there is also no reference group or peer arbiter of what is “good” or “bad” systemic program evaluation.

NSF therefore welcomes inputs from various sources, indeed often pays for them. Nonetheless, this is no guarantee that the evaluator will do a better or more thorough job of capturing the process of reform or specific measures of outcomes than site- or agency-based staff. The presentation at the NISE Forum will elaborate on this theme using NSF's **USI** program as an example.

⁸ A report on the NSB series of hearings, *Education on the Road to Excellence*, is in preparation. For agendas, participant lists, and highlights, see www.nsf.gov/nsb/committees.

VALUE-ADDED INDICATORS

Robert H. Meyer
Wisconsin Center for Education Research and
Harris Graduate School of Public Policy Studies
The University of Chicago

Introduction

Educational outcome indicators increasingly are being used to assess the efficacy of schools, programs, and policies. This paper discusses the weaknesses of the most commonly used educational indicators—average and median test scores—and the advantages of value-added indicators.⁵ Several major conclusions emerge from the analysis.

First, the typical indicators used to assess school performance—average and median test scores—are highly flawed as measures of school performance, even if they are derived from highly valid assessments. As a result, they are of limited value, if not useless, for evaluating relative school performance or school performance over time and thus should not be used to hold schools accountable for their performance. Indeed, simulation results indicate that changes over time in average test scores could very well be negatively correlated with actual changes in school performance.

Second, the typical indicators used to assess school performance are likely to provide schools with the perverse incentive to "cream," that is, to raise measured school performance by educating only those students that tend to have high test scores. The potential for creaming is apt to be particularly strong in environments characterized by selective admissions. However, creaming could also exist in more subtle, but no less harmful, forms. For example, schools could create an environment that is relatively unsupportive for potential dropouts, academically disadvantaged students, and

special education students, thereby encouraging these students to drop out of or transfer to another school. Second, schools could aggressively retain students at given grade levels. Finally, high quality teachers and administrators could gravitate to neighborhood schools that predominantly serve high-scoring students.

Third, typical school performance indicators tend to be biased against schools that disproportionately serve academically disadvantaged students. One source of bias is the well-known fact that school productivity is only one of the many determinants of student achievement. Most of the variation in average or median test scores can usually be accounted for by differences across schools in the types of students enrolled.

Finally, given the problem of student mobility (as well as several other problems discussed in this paper), it is not possible to construct statistically valid school performance indicators if tests, assessments, or other measures of student outcomes are conducted so infrequently that a significant proportion of students change schools in between periods of testing.

Given the substantial problems that exist with common educator indicators, what should be done to improve the situation? If one is interested in the use of indicators that are appropriate for accountability and/or evaluation purposes, I believe that the only solution is to design indicator/evaluation systems based on the value-added approach, as has recently been done in a number of districts and states—Dallas, Minneapolis, South Carolina, and Tennessee.

The essence of the value-added approach is that school performance is measured using a statistical regression model that includes, to the extent possible, all of the non-school

⁵ Many of the issues discussed in this paper are considered at greater length in Meyer (1996).

factors that contribute to growth in student achievement-in particular, student, family, and neighborhood characteristics. The key idea is to statistically isolate the contribution of schools to growth in student achievement over a given time period from all other sources of student achievement growth. This is particularly important in light of the fact that differences in student and family characteristics account for far more of the variation in student achievement than school-related factors. Failure to account for differences across schools in student, family, and community characteristics could result in highly contaminated indicators of school performance. A technical appendix summarizes these aspects of the value-added approach.

The next section presents a detailed critique of the average test score. The subsequent section considers the data requirements for value-added indicators.

A Critique of the Average Test Score as a Measure of School Performance

A school-level average test score is a highly flawed measure of school performance for four basic reasons. One, the average test score is contaminated by factors other than school-performance, in particular, the average level of student achievement prior to entering first grade-average initial achievement-and the average effects of student, family, and community characteristics on student achievement growth from first grade through the grade in which students are tested. In fact, it is quite likely that comparisons of average test scores across schools primarily reflect these differences rather than genuine differences in intrinsic school performance. As such, average test scores are highly biased against schools that disproportionately serve academically disadvantaged students and communities.

Two, the average test score reflects information about school performance that tends to be grossly out of date. For example, consider the average test score for a group of tenth grade students. The test scores for these students reflect learning that occurred in

kindergarten, roughly ten and a half years earlier, through the tenth grade. Indeed, a tenth grade level indicator could be dominated by information that is five or more years old. The fact that average test scores reflect **out-of-date** information severely weakens them as instruments of public **accountability**. In order to allow educators to react in a timely and responsible fashion, performance indicators presumably must reflect information that is current.

Three, average test scores at the school, district, and state levels tend to be highly contaminated due to student mobility in and out of different schools. For example, the typical high school student is likely to attend several different schools over the period spanning kindergarten through grade 12. For these students, a test score reflects the contributions of more than one, and possibly many, different schools. The problem of contamination is compounded by the fact that rates of student mobility tend to differ dramatically across schools. Contamination is apt to be especially high in communities that undergo rapid population growth or decline and in communities that experience significant changes in their occupational and industrial structure. Contamination due to student mobility is probably a relatively minor problem at the national level, since rates of in- and out-migration are low compared to rates of mobility within the nation; but at the state, district, and school levels, it is apt to be quite serious.

Finally, the average test score fails to localize school performance to a specific classroom or grade level-the natural unit of accountability in a traditional school. This lack of localization is, of course, most severe at the highest grade levels. A performance indicator that fails to localize school performance to a specific grade level or classroom is likely to be a relatively weak instrument of public accountability.

A Simulation

A simulation demonstrates vividly how the average test score is affected by past variation in school performance and hence is

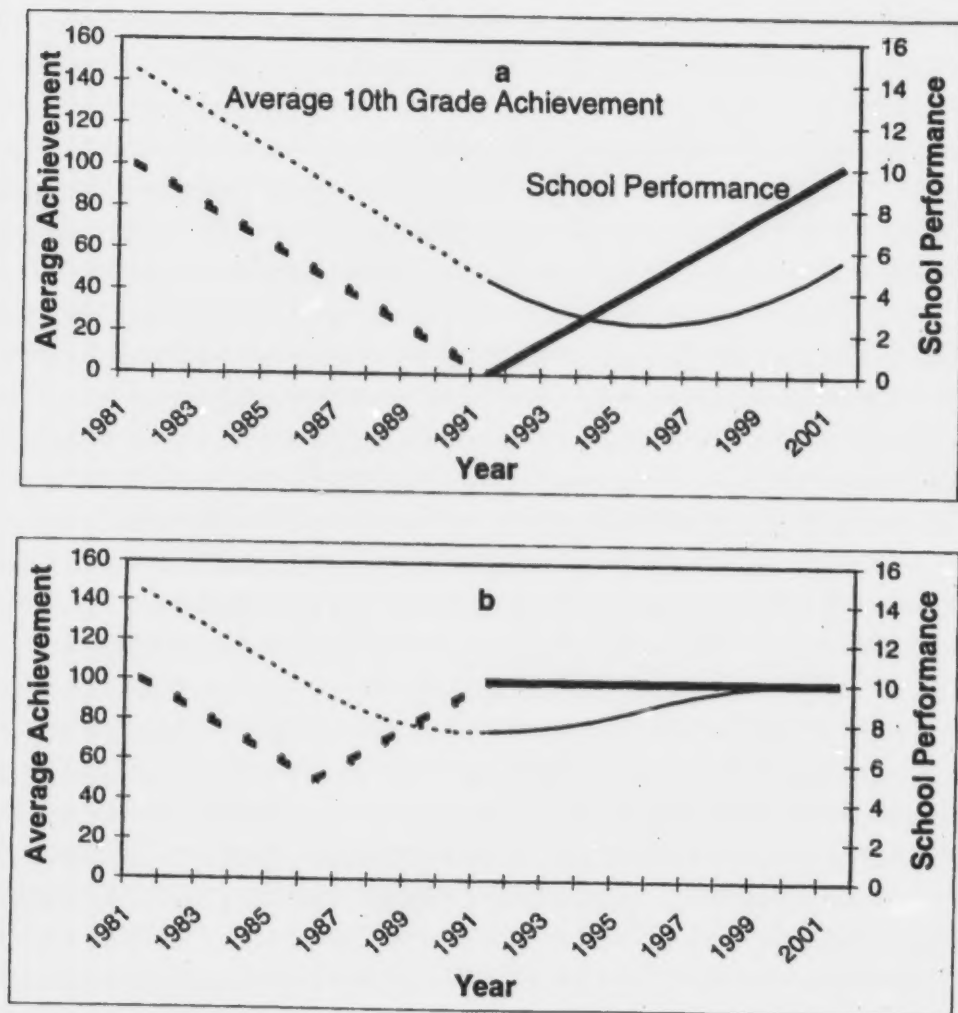
apt to be quite misleading as an indicator of current school performance.

To focus on the consequences of variation over time in school performance, assume that (1) average initial achievement and average student characteristics are identical for all schools at all points in time; (2) there is no student mobility; (3) the value-added school performance is identical at all grade levels in a given year; and, (4) growth in student achievement follows a linear growth process.

Given these assumptions, average student achievement at the end of a given grade at a given point in time is simply equal to the sum of value-added school performance indicators for all prior grades (beginning with first grade) in the appropriate prior years.

Figure 1 charts average tenth-grade achievement and school performance at a given point in time before and after the introduction of hypothetical academic reforms in 1992. This analysis is particularly

Figure 1. Average 10th Grade Achievement versus School Performance



Source: Data simulated by author.

relevant for the purposes of evaluating the efficacy of school reform efforts. Figure 1(a) depicts a scenario in which academic reforms reverse a trend of gradual deterioration in school performance across all grades and initiate a trend of gradual improvement in school performance across all grades. Figure 1(b) depicts a scenario in which academic reforms have absolutely no effect on school performance. The reforms, however, are preceded by an era of gradual deterioration in school performance across all grades, followed by a brief period (1987-1991) of gradual improvement across all grades.

Figure 1 illustrates that the average tenth-grade test score provides a totally misleading view of the effectiveness of the hypothetical academic reforms implemented in 1992. In Figure 1(a) the average tenth-grade test score declines for five years after the introduction of successful reforms. In 1(b), the average tenth-grade test score increases for a decade after the introduction of reforms that have no effect on student achievement growth. These results are admittedly somewhat counterintuitive. They arise from the fact that tenth-grade achievement is the product of gains in achievement accumulated over a ten-year period. The average tenth-grade test score is, in fact, exactly equal to a ten-year moving average of school performance. This stems from the simple assumption in this simulation that school performance is identical at different grade levels in the same year. The noise introduced by this type of aggregation is inevitable if school performance is at all variable over time.

An Example Based on National Data

The practical significance of the above analysis is illustrated using data on average mathematics scores from 1973 to 1986 from the National Assessment of Education Progress (NAEP). As indicated in Panel A of Table 1, NAEP scores for grade 11 exhibit the by now familiar pattern of sharp declines from 1973 to 1982 and then partial recovery between 1982 to 1986. The 11th grade data, by themselves, are fully consistent with the premise that academic reforms in the early

and mid-1980s generated substantial gains in academic achievement. In fact, an analysis of the data based on a gain indicator (a value-added type indicator) rather than an average test score suggests the opposite conclusion (see Panel B of Table 1).

The gain indicator is similar to a true value-added indicator in that it controls for differences among students in prior achievement. It does so in a very simple and intuitive way: gain is the change in average test scores over time (and across grades) for the same cohort of students. For example, the gain in test scores for students who were 11th grade students in 1986 is given by average test score of 11th grade students in 1986 minus the average test score for those students when they were 7th graders in 1982 (four grades and four years earlier); that is, $302.0 - 268.6 = 33.4$. Unfortunately, the gain indicator, unlike the value-added indicator, does not control for differences in student, family, and neighborhood characteristics that contribute to growth in student achievement. As a result, the gain indicator reflects possible changes over time in the composition of the population as well as changes in school productivity.⁶ Nonetheless, it is instructive to compare the gains in achievement experienced by different cohorts.⁷

As indicated in Panel B, the achievement growth of high school students (from 7th to 11th grade) during the 1982 and 1986 period was actually no better than achievement growth during previous periods. In fact, the gain from 7th to 11th grade was actually slightly lower during the 1982 to 1986 period than in previous periods! The rise in 11th grade math scores from 1982 to 1986

⁶ The gain indicator also cannot be constructed if the before (pre) and after (post) tests differ and have not been placed on the same measuring scale.

⁷ NAEP was originally designed to permit this type of analysis. In mathematics, the tests have generally been given every four years at grade levels spaced four years apart. For this illustrative analysis, we assume that average test scores in 1973 are comparable to the unknown 1974 scores.

TABLE 1
NAEP Mathematics Examination Data
(A)

Average Test Scores by Year				
Grade/Age	1973	1978	1982	1986
3rd/9	219.1	218.6	219.0	221.7
7th/13	266.0	264.1	268.6	269.0
11th/17	304.4	300.4	298.5	302.0

(B)

Average Test Score Gain From Year to Year			
Grade/Age	73 to 78	78 to 82	82 to 86
3rd to 7th/9 to 13	45.0	50.0	50.0
7th to 11th/13 to 17	34.4	34.4	33.4

Source: Dossey et al. (1988).

stems from an earlier increase in achievement growth for that cohort rather than from an increase in achievement growth over grades 7 to 11. In short, these data provide no support for the notion that high school academic reforms generated significant increases in test scores during the mid-1980s. These data also vividly confirm the general superiority of the gain indicator relative to level indicators, such as the average test score, as a measure of educational productivity.

It would be interesting to report the above analysis using true value-added, as opposed to gain, indicators. Unfortunately, the NAEP data do not permit such an analysis to be conducted, since the same students are not sampled for two consecutive NAEP surveys. This weakness in NAEP data could be

remedied by switching to a survey design that was at least partially longitudinal.

Value-Added Indicators: Data Requirements

Given the problems that exist with the average test score and other level indicators and, to a lesser degree, the gain indicator, it is important to consider whether value-added indicators could potentially be used as the foundation for school district, state, and national performance indicator/accountability systems. There are at least two reasons to be optimistic in this regard. First, value-added models have been used extensively over the last three decades by evaluators and other researchers interested in education and training programs. Second, a number of

districts and states, including Dallas, Minneapolis, and Tennessee, have successfully implemented value-added indicator systems.⁸

Nonetheless, despite the promise of value-added indicator systems, it is clear that they require a major commitment on the part of districts and states. In particular, districts and states must be prepared to: (1) assess students frequently and (2) develop comprehensive district or state data systems that contain information on student test scores and student, family, and community characteristics.⁹ The need for frequent testing stems from the fact that value-added indicators are designed to measure the contribution of schools to growth in student achievement over a given time period. In order to be able to construct value-added (or gain) indicators, it is therefore necessary to have achievement data for the same individuals at two points in time. Students who are missing either pre or post-test data must be excluded from the analysis and thus from a district's accountability system.

From the perspective of measuring school performance, an ideal testing program would do the following:

- Test all students annually during the late spring. Many districts currently follow this practice.
- Test all students who attend summer school at the end of the summer (or in the fall at the beginning of the subsequent school year). Following the recent boom in summer school enrollments, many districts have begun testing students at the end of summer school.
- Test mobile students at the point of entry into the district (or into a new school in the district).¹⁰ Minneapolis

⁸ See, for example, Clotfelter and Ladd (1996), Mandeville (1994), Millman (1997), Olson (1998), and Sanders and Horn (1994).

⁹ The latter data are required as control variables in the value-added model.

¹⁰ In principle, mobile students could also be tested prior to migrating out of a school or district. On

is one of the districts that is pioneering the use of entry-point testing. As indicated below, this is a very important component of a comprehensive assessment program. (Optionally, all students—including non-summer school students—could be tested in late spring and early fall. This would, of course, substantially raise the costs of testing.)

Annual testing has three major advantages. First, it maximizes accountability by localizing school performance to the most natural unit of accountability: the grade level or classroom. Second, it yields up-to-date information on school performance. Third, it severely limits the number of students who would be excluded due to student mobility and, as a result, yields a data set that is likely to be highly representative of the school population as a whole and large enough to yield statistically reliable school performance estimates.¹¹ Adding in a post-summer school test yields one additional advantage; namely, it allows districts to separately evaluate the productivity of schools during the regular school year and during the summer. Finally, adding a point-of-entry test for in-migrating students enables districts to include in-migrating students in their indicator systems. In an era where schools are increasingly under pressure to achieve high (measured) performance, it seems particularly unwise to

the other hand, these students might not have much of an incentive to take a test just prior to leaving a school and if they did take such a test, the results could be quite misleading. I do not see an easy way of including out-migrants in an accountability system other than testing all students at multiple points during the school year—an extremely expensive proposition.

¹¹ On the other hand, less frequent testing, such as testing at grades kindergarten, 4, 8, and 12, might be acceptable for national purposes, since student mobility is not really an issue at the national level. For purposes of evaluating local school performance, however, the problems created by student mobility argue strongly for frequent testing.

adopt an indicator system that systematically excludes any group in the population.¹²

One of the primary obstacles to developing a comprehensive data system is, in my opinion, the difficulty of collecting extensive information on student and family characteristics. This issue is potentially quite important because value-added indicators are often implemented using the rather limited administrative data that is commonly available in schools—for example, race and ethnicity, gender, special education status, limited English proficiency (LEP) status, eligibility for free or reduced-price lunch, and whether a family receives welfare benefits. Researchers equipped with more extensive data have demonstrated that parental education and income, family attitudes toward education, and other variables are also powerful determinants of student achievement growth.

The consequence of failing to control adequately for these and other student, family, and community characteristics is that real-world, value-added indicators are apt to be biased because they absorb differences across schools in average unmeasured student, family, and community characteristics as well as differences in intrinsic school performance. This implies that a value-added indicator derived from a model with "weak" predictors of student achievement growth might be only slightly better than a gain indicator (better in the sense of being more highly correlated with

a theoretically perfect value-added indicator). Even so, it is likely to be a much better indicator than the average test score.

The key issue, of course, is not whether a feasible value-added indicator is perfect. Rather, the issue is whether the indicator provides a substantially better measure of school performance than other affordable indicators and whether it is good enough so that, on balance, it makes a net, positive contribution to the school improvement process, relative to other possible indicator/school accountability systems.

Conclusions and Recommendations

The average test score, one of the most commonly used indicators in American education, is highly suspect as an indicator of school performance.¹³ This indicator suffers from four major deficiencies: it fails to localize school performance to the classroom or grade level; it aggregates information on school performance that tends to be grossly out of date; it is contaminated by student mobility; and, it fails to measure the distinct contribution of schools to growth in student achievement, as opposed to the contribution due to students, families, and community factors. As a result, the average test score is a weak, if not counterproductive, instrument of public accountability. The gain indicator, if it can be computed, and the value-added indicator avoid three of the four problems that plague the average test score. The feasible value-added indicator has the major advantage that it potentially eliminates the bias that exists in the gain indicator due to differences across schools in student, family, and community characteristics, particularly if it is based on a model that includes an extensive set of control variables. In this case, it fully eliminates the incentive for schools to cream.

The value-added approach to measuring school performance relies on a statistical model to identify the distinct contributions made by schools to growth in student achievement. The quality of a value-added

¹² To limit the costs and burden imposed by frequent student tests, it might be feasible to vary the frequency of testing across schools. Provided that a district tests all in-migrating students, annual testing could be implemented only in schools where student mobility is high. In addition, annual testing could be implemented in areas with limited enrollments in order to improve the reliability of estimates in these areas and in schools with low measured performance in order to monitor these schools with greater vigilance. In all other schools, students could be tested in every other grade. One major disadvantage of this approach is that performance indicators for the latter set of schools would reflect the performance of teachers at two different grade levels rather than at a single grade level.

¹³ Other level indicators, such as the median test score, are similarly suspect.

indicator is determined by four factors: the frequency with which students are tested; the quality and appropriateness of the tests that underlie the indicators; the adequacy of the control variables included in the appropriate statistical models; and, the technical validity of the statistical models used to construct the indicators.

In terms of the first issue, I believe that states and districts need to seriously consider testing students at every grade level, beginning with kindergarten. To further improve their indicator systems, states and districts need to think about testing summer school students and in-migrating students at the point of entry into the school or district. With respect to the second and third issues, it is important that states make it a major priority to collect extensive and reliable information on student and family characteristics and to develop state tests that are technically sound and fully attuned to their educational goals. Finally, further research is needed to assess the sensitivity of estimates of school performance indicators to alternative statistical models.

Technical Appendix: A Simple Value-Added Model

In order to evaluate the validity of alternative school performance indicators it is necessary to specify the "standard" that will be used to evaluate all other indicators. Consistent with the vast literature on the determinants of achievement growth, I define true school performance using a statistical, value-added model (see, for example, Dyer, Linn, & Patton, 1969; Hanushek, 1972; Hanushek & Taylor, 1990; Meyer, 1994, 1996; Millman, 1997; Murnane, 1975; Raudenbush & Willms, 1995; Sanders & Horn, 1994; and, Willms & Raudenbush, 1989). In order to simplify the presentation, I assume that achievement growth from one grade to the next can be adequately characterized by the following simple value-added model: a two-level, linear growth model.

$$Y(i, g, t) = Y(i, g - 1, t - 1) + \beta(g)X(i)$$

$$+ \sum_s [\alpha(s, g, t) + \gamma(g)C(s, g, t)] S(i, s, g, t) + e(i, g, t)$$

where i indexes individual students, s indexes schools, g indexes grade levels, and t indexes school years; Y represents student achievement for a given individual in grade g in year t ; $X(i)$ represents a set (vector) of individual and family characteristics, assumed (for simplicity) to be invariant over time; $S(i, s, g, t)$ is an indicator variable equal to one if student i is enrolled in schools s in grade g in year t , zero otherwise; $C(s, g, t)$ represents a set (vector) of community characteristics and school-aggregate student characteristics (for example, average school socioeconomic status); $\beta(g)$ and $\gamma(g)$ are parameters (vectors) that capture the effects of X and C on growth in student achievement; $\alpha(s, g, t)$ is a school effect; and $e(i, g, t)$ is a random component of student achievement growth assumed to be uncorrelated with all regressions included in the model.¹⁴

Despite the notational complexity of the model, it has a straightforward interpretation: student achievement in a given grade in a given year is equal to student achievement in the prior year and grade, plus a term $\beta(g)X(i)$ that reflects the contribution of individual and family characteristics to growth in student

¹⁴ The above model only makes sense if the pre- and post-test scores are scaled so that achievement is measured in the same units. If this is not the case, the model could be extended to allow the pre-test variable to have its own coefficient, possibly different from the value of one that is imposed in the above model. In fact, this model has often been used in previous studies. However, Meyer (1992) demonstrates that in a model of this type it is necessary to correct for measurement error in the pre-test variable. Also note that the model is defined only for students who attend a given school for the entire school year and have achievement test data both prior to and at the end of the school year. Students who fail to meet these conditions must be excluded from the analysis. In principle, this problem could be avoided by testing students more than once a year, although this would be an expensive and burdensome proposition.

achievement, a term (in brackets on the second line) that reflects the growth in student achievement as a result of attending a given school, and a random term. The effect of attending a given school can further be broken down into two parts: a component $\alpha(s,g,t)$ that is the result of differences in school policies, teacher quality, etc., and the contribution of community and school-aggregate student characteristics.

The former factor $\alpha(s,g,t)$ measures the contribution of a school to growth in student achievement after controlling for all factors that are external to the school.¹⁵ I refer to this indicator as a measure of intrinsic school performance.¹⁶ Willms and Raudenbush (1989) refer to this indicator as a Type B indicator. This indicator can be interpreted as a measure of the collective performance of school staff (at a given grade level) and thus is the indicator that is appropriate for purposes of school accountability. A second value-added indicator, a measure of total school performance, is given by $\alpha(s,g,t) + \gamma(g)C(s,g,t)$.¹⁷ Willms and Raudenbush refer to this indicator as a Type A indicator. This indicator reflects the intrinsic performance of a school (α) plus the part of school performance that is determined by factors

¹⁵ In practice, this indicator is contaminated somewhat by the random error $e(i,g,t)$. For each school, $\alpha(s,g,t)$ absorbs the average of all student errors. Since these errors are zero, on average, the estimated value of $\alpha(s,g,t)$ is unbiased.

¹⁶ The intrinsic school performance indicator is implicitly defined by the school-level control variables (C) included in the model. At a minimum, school-level measures of student and neighborhood characteristics should be included in the model since these variables are determined externally to the school (at least in the short term). The model could also include variables such as per pupil expenditures and the quality of building facilities in order to control for school inputs that may not be controllable by principals and teachers.

¹⁷ The total performance indicator can be obtained directly as the (fixed effects) coefficient on a school indicator in a model that excludes school-level variables (C). Bryk and Raudenbush (1989) and Meyer (1996) discuss methods for estimating the intrinsic school performance indicator (α).

external to the school; for example, community and school-aggregate student characteristics.¹⁸ One interpretation of this indicator is that it captures the effect of enrolling one additional student in a school, holding community characteristics and the composition of the student group approximately fixed. Thus, it is appropriate for purposes of informing school choice. In practice, the two indicators may yield quite similar rankings of schools. If so, it may be perfectly acceptable to report results from only one of the indicators. The total performance indicator is typically easier to compute.

The value-added indicators derived from equation (1) define school performance at a specific grade level at a particular point in time. The average test score, on the other hand, reflects the cumulative contribution of school, family, and community inputs in all grades prior to the year in which the students are tested. As indicated in the text, there are likely to be substantial differences between the two types of indicators.

References

- Bryk, A. S., & Raudenbush, S. W. (1989). Quantitative models for estimating teacher and school effectiveness. In R. D. Bock (Eds.), *Multilevel analysis of educational data*. San Diego: Academic Press, Inc., pp. 205-232.
- Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable*. Washington, DC: The Brookings Institution, pp. 23-63.
- Dossey, J. A., Mullis, I. V., Lindquist, M. M., & Chambers, D. L. (1988). *The Mathematics Report Card: Are we measuring up?* Princeton: Educational Testing Service.
- Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of

¹⁸ See Gamoran (1992) for a survey of research on the effects of school-aggregate student characteristics on school performance.

- obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal*, 6 (4), 591-605.
- Gamoran, A. (1992). Social factors in education. In M. Alkins (Ed.), *Encyclopedia of educational research* (6th edition). New York: Macmillan, pp. 1222-1229.
- Hanushek, E. A. (1972). *Education and race*. Lexington, Mass.: D.C. Heath.
- Hanushek, E. A., & Taylor, L. (1990). Alternative assessments of the performance of schools. *Journal of Human Resources*, 26 (2), 179-201.
- Mandeville, G. K. (1994). The South Carolina experience with incentives. In T. A. Downes & W. A. Testa (Eds.), *Midwest approaches to school reform*. Proceedings of a Conference held at the Federal Reserve Bank of Chicago, October 26-27, pp. 69-97.
- Meyer, R. H. (1992). *Applied versus traditional mathematics: New econometric models of the contribution of high school courses to mathematics proficiency*. Discussion Paper No. 966-92. Madison: University of Wisconsin-Madison, Institute for Research on Poverty.
- Meyer, R. H. (1994). *Educational performance indicators: A critique*. Discussion Paper No. 1052-94. Madison: University of Wisconsin-Madison, Institute for Research on Poverty.
- Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives*. Washington, DC: National Academy Press, pp.197-223.
- Millman, J. (1997). *Grading teachers, grading schools*. Thousand Oaks, CA: Corwin Press, Inc.
- Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger Publishing Co.
- Olson, L. (1998). A question of value. *Education Week*, May 13, pp. 27, 30-31.
- Raudenbush, S. W., & Willms, D. J. (1991). *Schools, classrooms, and pupils*. San Diego: Academic Press.
- Raudenbush, S. W., & Willms, D. J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20 (4), (Winter), 307-336.
- Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8, 299-311.
- Willms, D. J., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26, 209-232.

QUANTITATIVE AND QUALITATIVE DATA IN THE THEORY OF SYSTEMIC REFORM

William H. Clune
Wisconsin Center for Education Research

In a recent NISE paper, I applied a theory of systemic reform to nine case studies of State Systemic Initiatives (SSIs) published by SRI in March, 1998 (Clune, forthcoming). This theory was developed by the policy studies team of NISE (Eric Osthoff, Paula White, and myself) from many sources of information, including workshops, Forums (like this one), documents, and interviews of SSI staff. In this paper, I will focus on one main methodological aspect of the longer paper that is relevant to the theme of

evaluation in this Forum: the usefulness and limits of quantitative ratings as a tool for understanding systemic reform.

Cross-Site Ratings Based on the Theory

To start the discussion of quantitative ratings, consider the following table from the larger paper, which rates the nine SSIs according to four variables, each measured by breadth and depth:

Table 1
Breadth, Depth, and Average Ratings of the Nine SRI States

STATE	REFORM		POLICY		CURRIC.		ACHIEVE.		AVG.
	Br.	Depth	Br.	Depth	Br.	Depth	Br.	Depth	
Connecticut	4	4	4	4	3	2	4	4	3.6
Maine	4	4	4	4	3	2	4	4	3.6
Montana	3	4	2	4	2	3	2	4	3.0
Louisiana	4	4	3	2	3	2	2	2	2.8
Michigan	2	3	2	2	2	2	3	2	2.3
California	2	3	2	3	3	2	2	1	2.3
Arkansas	3	3	2	2	2	1	2	2	2.1
Delaware	2	1	1	1	1	1	1	1	1.1
New York	1	1	1	1	1	1	1	1	1.0
AVERAGE COLUMN RATINGS	2.8	3.0	2.3	2.6	2.2	1.8	2.3	2.3	

First, a few words of explanation about what Table 1 means. Our theory is that systemic reform is a continuous cycle of causation running from left to right across the column headings: reform, policy, curriculum, and achievement. This sequence is faithful both to the theoretical origins of systemic reform and our observations of the SSIs; it is inductively derived and deductively grounded.

Theoretical origins include the hypothesized link between systemic policy and major upgrading of the curriculum (Resnick & Resnick, 1985; Smith & O'Day, 1993), as well as the hypothesized link between curriculum and achievement, sometimes called opportunity to learn (Bryk, Lee, & Smith, 1990). Deductive theory-building was especially useful for the reform variable,

because it was not clear in advance how the SSIs would go about trying to change the system.

In the next row in the Table, each of the four variables is rated according to depth and breadth. Breadth refers to what Norman Webb's group calls "saturation," the proportion of the system that has been changed in a standards-based direction. Depth refers to the strength of the influence of reform and policy, or the intensity of the change in curriculum and achievement. Both breadth and depth are essential aspects of a theory of systemic reform. The breadth hypothesis is that alignment across a wide range of policies will produce changes in many schools. The depth hypothesis is essential to a causal theory of systemic reform: Given any amount of breadth, more change in one variable will be produced by greater strength in the preceding variable.

The Table does not show the detailed definitions (or operationalizations) of the four variables that guided the rating exercise. For each variable, we developed a list of detailed attributes defining depth and breadth; and it is these specific characteristics that were tested against the data. For example, the attributes of successful reform (or the tools of reform) are: vision, strategic planning, networking with policy makers, networking with professionals, institutionalization of the reform structure, leveraging of resources, and public outreach and visibility. The reform would be considered broad to the extent it had all of these elements, and the elements touched all the levers of policy, and deep to the extent that each element was well developed and predictively influential.

Breadth and depth were each rated on a five-point scale, producing the numbers in the table. You see that the actual results do not include any fives and do include many ones and twos, a fact that will be analyzed below.

The Usefulness of Quantitative Data in Testing the Theory

An important first question about Table 1 is why reduce all of the complexity of the SSIs to a set of numbers in one table, and the

answer is that it provides both a means of testing the theory and of reaching certain conclusions about how reform works.

To test the theory of systemic reform, we must be able to say whether higher levels of systemic reform and policy do, in fact, produce more change in teaching and learning. The numbers in the table provide some support for the theory. Higher ratings in reform and policy generally are associated with broader and deeper change in curriculum and achievement. On the other hand, we can also say from the many low ratings that as a group the SSIs fell far short of achieving the ultimate goal of transforming entire states. In this group of SSIs, no state achieved the goal of complete transformation, several states were just beginning, and we do not expect higher results in our larger study of all of the SSIs. The usual explanations of partial success are the short duration and low funding of the reforms, explanations for which we find support in qualitative evidence, discussed below. The main point here, however, is simply the usefulness of numbers in answering fundamental questions about the validity of the theory, or in other words, about the success of reform.

The Usefulness of Quantitative Data in Making Generalizations About Reform

Table 1 also can be used to reach a number of important generalizations about systemic reform, for example, by looking at the average ratings of each variable across the states (bottom row in the Table). Reform and policy are stronger than curriculum and achievement. Greater strength in reform and policy might be expected both because of the sequence of reform (with those areas receiving attention first) and the sheer difficulty of making an impact on teachers and students. The lowest column rating is for depth of influence in the curriculum, and exactly this-shallow influence on the curriculum-was identified as the chief failing of systemic reform in an earlier research synthesis sponsored by NISE (Knapp, 1997). The significance of this low rating will be discussed next.

Interpreting the Quantitative Data from Qualitative Data

Numbers cannot tell us everything we need to know, a first demonstration of which is the need for qualitative information as a means of interpreting the numbers. Consider, for example, the low rating of the curriculum column on the table, indicating generally low breadth and depth in measured changes in course content and pedagogy. Detailed familiarity with the case studies and other data led us to two conclusions about the reasons for the low ratings: first, substantive problems with choosing policies with a relatively weak influence on the classroom; and, second, problems of measurement and research design.

A common substantive problem is the pedagogical orientation of reform—its emphasis not simply on teacher training but on active learning as a pedagogical technique. Reforms typically were aimed at classroom processes such as the use of manipulatives, collaborative learning, and inquiry learning. Especially early in the reforms, direct means of influencing curriculum were relatively rare. It is surprising how few reforms focussed on what the students were being taught as opposed to how. The gap between pedagogy and content narrowed as the reforms progressed, partly as a result of prodding by NSF. By the mid-90s, many of the stronger reforms were using new materials, model teaching units, or curriculum replacement units.

Few of the SSIs began with any systematic observation of changes in course content and pedagogy embedded in a research design that would allow comparison of change in schools or time periods more and

less affected by reform. The best evidence of curriculum change that is available on many SSIs is data on the number of teachers trained, surveys of teacher attitudes and practice, and evidence of some whole-school restructuring. The absence of good data is a problem not just for measurement of curriculum change but for testing the theory as a whole. Any theory or evaluation of systemic reform requires testing causal links in complex systems on the basis of relatively few cases (observations). The task of theory testing would be much easier and the case much more convincing if there were more direct and precise data on teaching and learning that could be associated with varying degrees and phases of reform. States were certainly moving in that direction—for example, evaluations that compared gains in student achievement with the number of SSI-trained teachers in schools; but the effort is truly in its infancy. Here we see a possible link between methodology and the substance of policy design. A main reason for the lack of early data may be that classroom change was not a clear objective of policy. Many SSIs were built around teacher enhancement projects (professional development), where teacher capacity is the goal rather than upgrading the curriculum.

Generalizations Beyond Quantitative Data

The limited usefulness of numbers also can be seen in the importance of many generalizations about systemic reform that require purely qualitative analysis, such as synthesizing patterns across sites. For example, my paper reached the following conclusions about the characteristics of reform in the more successful of the nine states:

Table 2
Common Characteristics of Successful Reform

- An independent agency well connected to higher education
- Working in the middle of policies and other institutions
- A strong state assessment
- Teacher and professional networks connecting policy and schools
- Incremental building on earlier periods of reform

The paper also reached the following conclusions about common deficiencies of systemic reform across sites, two of which I

have already mentioned as interpretations of the quantitative data.

Table 3
Common Deficiencies of Systemic Reform

- The absence or indirection of influence over curriculum content
- The dearth of fully aligned state assessments
- The absence of good data on classrooms and student achievement
- The slow growth of whole-school restructuring
- The unexplored territory of adequacy in urban schools

Conclusion

In conclusion, this paper has argued that numerical ratings are highly useful in building and testing a theory of systemic reform. No other technique seems as capable of testing the basic hypothesis that strength in one variable produces strength in the next, summarizing the overall progress of individual reforms and the reform movement, analyzing the status of particular components of reform across different sites (e.g., curriculum change), and raising important questions for interpretive qualitative analysis. At the same time, quantitative data are no substitute for careful qualitative analysis and synthesis. First, many of the data that go into the quantitative analysis are themselves the result of qualitative inquiries, such as interviews (e.g., the nature of the reformers' vision). Second, many important institutional patterns across reforms can only be recognized and understood by means of thoughtful qualitative inquiry.

References

- Bryk, A., Lee, V., & Smith, J. (1990). High school organization and its effects on teachers and students: An interpretive summary of the research. In W. H. Clune & J. F. Witte (Eds.), *Choice and Control in American Education: Vol. I. The Theory of Choice and Control in Education*. London: The Falmer Press.
- Clune, W.H. (forthcoming). *Toward a theory of systemic reform* (Research Monograph). Madison, WI: University of Wisconsin, National Institute of Science Education.
- Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher*, 13 (4), 5-20.
- Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S.H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-267). Philadelphia: Falmer.

Introduction to Breakout Session III Question Summary

Each panel was followed by a Breakout Session. Participants were assigned to small groups of ten to twelve, led by a facilitator, in a discussion of three questions and other issues raised by the presenters. Each set of three questions was developed by the organizers of the Forum. At the beginning of the Breakout Session, participants were asked to write their responses to each of the three questions on index cards. The comments that the participants wrote were used to begin the small group discussions. The index cards were given to two people, who provided a synthesis of the conference; comments on the index cards were incorporated into their comments. Responses to the first question are summarized here to provide examples of participants' comments.

Breakout Session III: Findings on Systemic Reform from Evaluation and Research

Participants' Comments:

Q: How can effects and impact be attributed to systemic reform?

A total of 66 participants responded to this question. As was the case in response to the other questions, participants offered a range of views on attributing impact to systemic reform. Five of the responses (8%) either felt this was the wrong question to ask or did not think it was possible to attribute any effects to systemic reform. One participant responded, "I wonder about the pressure to claim attribution-to piece apart the general success. Even specific classrooms that are successful have a context and complexity that are both challenging and difficult to attribute claims." Another participant made a similar point, "Carefully. By its very nature, systemic reform efforts and their impact are messy and sometimes/often unpredictable. There are so many forces at work that direct effect is difficult to attribute [to any one thing]."

Another participant gave a qualified response:

Because the SI is generally a minor player in terms of resources brought to the change effort and the perceived influence relative to other factors pushing/pulling on the system, attribution will always be fuzzy at best. I agree with Mark St. John that student performance measures are the riskiest to attempt attribution; system capacity measures are probably more directly amenable, although not in a rigidly quantitative manner.

Most of the respondents suggested experimental designs or methodologies that they felt could help to attribute impact to systemic reform.

1. *Use quantitative and qualitative methods in combination.* Eight participants (12%) cited the use of both quantitative and qualitative methods. One

participant gave a cryptic response, "Careful combination of qualitative and quantitative measures—creative use of funding illuminated by case study information." Another participant was more explicit, "The analysis of the reform may accurately reflect the effects/impact when multiple approaches are employed: growth in achievement over a period of time, case studies, selection of students of similar achievement levels across a broad sampling and comparisons of growth."

2. *Employ baseline measures along with other measures to track change.* Seven participants (11%) suggested that the attribution question be answered by trying to measure change over time, which required having baseline information. One participant responded, "Change before/after, products-curricular, catalyst, profiles." Another participant reflected, "Seems to boil down to looking at trends and reversal of trends. Most presenters have underscored that attribution is really tough. Perhaps tracing cohorts' performance before and after the reform process (but the assessment may be different)." A third participant succinctly stated, "Compute achievement gains, both baseline and year-to-year."

3. *Compute the value-added by systemic reform.* Six participants (9%) indicated the importance of using value-added methodologies. "It seems like we shouldn't trust student achievement data if [they aren't] derived through value-added methodologies," wrote one participant, referring to the presentation by Meyer. Another participant recognized the timeframe required to detect impact, "Over a long period of time, effects lag behind initiative-i.e., changes in teacher practice and student achievement need to be communicated using the value-added technique." A third participant supported using value-added methodology, but cautioned that the technique may not be as meaningful to all of the audiences. This participant responded, "Some kind of

value-added approach is needed, but can become too sophisticated for most audiences very quickly."

4. *Impose a control study.* Six participants (9%) indicated that some form of control group should be used in the evaluation. Half of these felt that the control group could come from within the system, such as the model used in Ohio that was described by Kahle, or that the control group should be with another comparable system (state or district). One participant suggested, "Controlled studies comparing SSI states or other systemic projects and comparable non-SSI states (comparing elements of the systems and outcomes). Careful documentation of the system and how funding intervened and produced certain outcomes." Not all participants agreed that control studies are appropriate to attribute impact to systemic reform. One participant pointed to the difficulty in using this methodology along with its potential benefit, "It is very difficult to extrapolate from a control study to a whole system. Lessons learned from these controlled studies can be used to convince policy-makers and administrators to implement a full-scale effort."

Participants in the third breakout session gave a range of other responses on how impact or effects could be attributed to systemic reform, but none of these responses were given by more than three of the participants. Some participants felt it was necessary to delay any attempt to attribute impact until systemic reform had had the time to mature. Others suggested using specific techniques, such as logic models, **meta-analysis**, and establishing a chain of evidence. A few noted specific variables that should be included in the analysis, including process variables, schools, and indicators.

One participant was troubled that the focus of the attribution question nearly always attended to outputs only. This participant raised a number of questions that need to be considered in thinking about attributing impact, "I think Clune came the closest to attending to the fingerprints of the individual systems. If the issue is systemic, then attention must adhere to the parts. Where are the vantage points? What happens when you use them? How long does it take to find them? What cultural impacts are there? What indicators of progress make sense in this system? Was it really news to everyone else that achievement data are insensitive indicators of change?"

Synthesis: 1999 NISE Forum

Cora B. Marrett

Provost

University of Massachusetts-Amherst

Several years ago, the Honor Society, Phi Kappa Phi, produced a volume entitled, *Making Heroes Out of Scholars*. Systemic reform seems likely to make heroes out of evaluators. For what is more challenging than the matter of trying to evaluate systemic reform. It is my pleasure, then, to join with the heroes, those already made, and those who are emerging.

It seems to me that this Forum has made three very key points. One of them is that there is convergence around the problems that plague the evaluation of systemic reform. There is some level of agreement among the people addressing these issues. Thus, we are not going off in totally different directions, as we try our heroic approaches. The Forum has also pointed to encouraging signs suggesting that the problems that have been identified are being tackled. And tackled in innovative, engaging ways. Finally, consistent with what Ernie House has just said, the Forum has hinted at the importance of leadership. Leadership on the evaluation front is required if the promises to be fulfilled are in fact to be realized, and if we're going to move to having those on the front line appear heroic and not fool hardy. Let me begin then with a few comments about the task of heroes, of evaluators as heroes.

The difficulty most people talked about in several sessions seemed to center on the fact that we are dealing with something that is complex and dynamic. The break out groups returned time and again to the question of how in the world does one try to capture, try to evaluate, try to do assessments in a system, or systems, that seem to be changing and that it is very hard to get one's hands on. Now, interestingly, complexity and dynamism are not necessarily limiting factors. Consider the advances made in our efforts to grasp

complex and dynamic physical, biological, and other systems. Thus, it does not seem that, by itself, complexity or dynamism should limit what we can do. But perhaps a number of the problems, as people have identified and that Ernie has certainly gone into, in part, is the conceptualization of system. What do we, in fact, mean by system? As he has indicated, one sees some shifts over time and from one context to another. We are very fortunate to have Mike Smith here. Mike will talk about what he, early on, had meant in discussing a system and alignment of the components of the system. That is not always, as Ernie has already suggested, the way in which others have talked about system; but these different definitions-meanings of the term **system**—might be one of the reasons why it becomes difficult to talk about evaluations in a shared way. It could be, as well, that the reasons why we have different components and different ideas coming into play, is that we really do not understand very well those components that make up systems. Even when we think we are using the same terms to describe them, they are not necessarily going to be the same in concept as the system we are actually approaching. It could also be that some components that matter in one setting might not matter as much in another. This certainly came across in one of the break out groups when someone said, "Imagine the role that culture plays when one is talking about a rural systemic initiative in the southwest."

Maybe it's not as important as in other kinds of contexts, but these questions about the components of a system and what components will matter under what sorts of conditions are thorny issues, and probably are among the reasons why we do not always have uniform kinds of assessments, uniform evaluations of systems, and of systemic

reform. It is the case, and I'll return to this momentarily, that what is taking place is that our theoretical notions are evolving along with the evaluation process. As a number of people said, that's really not, at least in the academic world, the way you usually want things to develop. You usually want to have some understanding of how things go together and, then, judge how these work out in given conditions. But here we are doing a number of things simultaneously. The participants, in addition to talking about complexity and dynamism—these also got considerable attention, especially from those in the break out sessions—were talking about what people call the politics of evaluation. They were saying, this is not something that can be addressed apart from questions of trying to demonstrate positive outcomes for a variety of other purposes: purposes that are not always associated with questions about achievement. I do not think we spent, in the larger sessions, as much time on the issues of the politics as probably occurred in the smaller groups. We've got, then, political questions on the table as well as conceptual questions.

One might ask then, why, in the midst of such challenges, would there be the energy, and the excitement about evaluating systemic reform? Why in the world are people sitting here, feeling committed to such an activity? This is where I want to turn to some of the positive sides, but let me get one of the items off the table immediately. It's not the money. It is clear that the activities are underfunded, that the research needed to advance our knowledge is not being supported at such high levels that we are going to see all of the problems solved in a short time. In fact, Iris Weiss and Zoe Barley, among others, drew our attention to the inadequacies of support for building the knowledge base. We are not seeing people involved here because we've got the future of a Bill Gates of evaluation on systemic reform. Instead, it seems that there are other reasons why people are committed and excited. One of these is the sense that there is a shared commitment to understanding things systemically. In at least one break out group, there was recognition that we have passed a threshold. We no longer

have to ask, is there something out there, whether we call it systemicness, or whatever the term we use. We have passed that threshold and reached the point at which there is an appreciation, at least some recognition, of what matters. Components do occur in some kind of relationship that can make a difference. The broader acceptance, then, of the idea of system seems at least to be enough to cause some people to say, "We're on the right track. We've got some other support for our ideas." This is also, it seems, the reason why people are very excited about what has taken place. There is evidence that linkages can be established and communicated. This came across extremely clear during the break out sessions. We heard, "I am encouraged by the findings from Puerto Rico." "I feel a lot better because Ohio has shown me that there are things that one can establish connections on." They went on to suggest that those kinds of experiences seem to indicate that the more closely aligned the assessment is with the changes in the system, the more convincing is the case. When there is a great gap between what you say are the outcomes and the changes that were made this is not a terribly convincing argument. In addition to the fact that there is evidence that the more convincingly the case is presented, the better is the response to what has in fact been uncovered. Then, there is the fact that a number of people have said, "What I've gained from this particular **forum** is that there are connections that can be made."

There were others, too, who were encouraged by the theoretical advances made and promised. Reading the papers, hearing the discussions from such people as Zoe Barley, Dan Heck, Bill Clune, Patrick Shields: All of these contributors suggest that there is a logic in the advances we are making. There are ways in which things connect. And it's that logic that must be unraveled and that must undergird the evaluation efforts. From the break out groups comes recognition that reform lies along a continuum. It's best viewed that way and not as a discrete phenomenon. It's the same kind of idea that was fundamental to the Clune **discussion**—that there are dimensions that can be specified

and relationships that can be better understood. Several of these dimensional continuous approaches seemed to be exactly what people were looking for as they struggled with what we can say about evaluation and systemic reform. More than one person expressed excitement about joining in the push for theory. And as I said, "Those of us from academia of course love that kind of view." Instead of saying, "We're at a disadvantage because we do not have all the underlying theory," they said, "We as evaluators are having a chance to advance the theory by our own participation in the development of knowledge." That was significant to a number of people and, in fact, it helped promote their excitement about this adventure.

There were those who said, "What I see [coming] from the Forum are some important methodological answers." Let me say something about what some people in the hallway said. They said, "You know, I've got answers to questions I never asked." This left some people a bit puzzled-not knowing some of the debates people are engaged in. But I came here hoping to get some answers to questions such as, How do I design the evaluation I'm trying to work on? How do I deal with the multi-level analysis that I've got to undertake? People said, "That's the kind of thing that I come away from the Forum with, having heard the discussions-having heard Heck's analysis, for example, and Meyer's analysis on value added-it makes some sense that there are particular answers to the thorny methodological design questions so many are confronting. The problems are, in fact, tractable.

There were those, too, who said, it might not have been talked about a lot, but there are signals from the Forum that even funders learn. If one looks at the evolution as discussed in the **Chubin** paper on NSF (National Science Foundation) over time, it is quite clear that the expectations noted at the beginning of the systemic reform activities, on the basis of their evaluations-those activities, those expectations-have changed. In many ways, this would seem a consequence of the feedback of the interaction

from the principal investigators, the sites, and, I would also say, from the evaluators. There is this idea that the evaluators participating in the process have important roles to play: as they have already played those roles in trying to do some of the things people suggested, helping often to unravel, helping in some cases to give voice to those kinds of concerns that exist at particular sites. Evaluators have the role of trying to clarify values and to indicate the different value assumptions that are brought to the table by those engaged in systemic reform. The evaluators, then, play an exciting role in so many ways. At this Forum, it certainly seemed as if they said, "Not only do we like what we have done, but we want to be able to continue this leadership role."

This brings me to the final point about the leadership that needs to be exercised. It seems people are requesting that some groups, maybe NISE, provide evaluators with even greater opportunities-opportunities to learn, opportunities to learn from one another about the experiences involved in evaluating systemic reform. Evaluators need opportunities to exchange ideas about the things that have worked, and under what conditions, in order to evolve the concepts, the methods, and the capacity to handle the kinds of things that are going to be important for moving the whole area of systemic reform ahead. Such opportunities, in fact, might transcend the areas of mathematics and science. Norman Webb reminds us that when it comes to reform, and site-level management, there are few districts, or schools, that can afford to separate the reform effort and its evaluation. There are the interactions that cut across subject areas, that demand the kind of attention, and, thus, the leadership that we're talking about. This is a leadership that thinks of reform in broad terms, that thinks about participation quite broadly, and that would not restrict the evaluation only to areas of interest to the source of funding. I pass on a comment that comes from one of the groups. A sail boat analogy. "Rather than build momentum to push through ill winds, build the capacity of the crew to set the sails in ways that maintain the course; assure that your crew can be

depended on to use good maps that show the shoals and the safe harbors." This is a somewhat lengthy analogy, but I think you get the idea about the crew, about the maps, and about avoiding the shoals.

Let me, in my final comments, applaud the heroic efforts of the evaluators to this point, and let me then join with others in hoping that you, the map makers, will continue to show the safe harbors for student achievement. Thank you.

Synthesis: 1999 NISE Forum

Ernest House
Professor
University of Colorado-Boulder

We have heard twelve papers and a lot of discussion in between. I would like at this point to try to make sense, or my sense at least, of what these papers mean, since they are all so different and present different proposals for the evaluation of systemic reform. Then I want to connect that to the concerns that I heard expressed in the **groups**—concerns that the participants have. Finally, I would like to make a few suggestions as to how the National Science Foundation (NSF), or some of its agents, might proceed in the future with evaluating systemic reform. I'll try to be brief.

First of all, what is systemic reform? This question came up time and time again in the groups. Smith and O'Day, in their definitive paper, make reasonably clear what, in their view, they think systemic reform should be—alignment of goals, standards, content, and assessment. This is not an entirely new idea. Ralph Tyler, who many years ago was considered the father of educational assessment, thought that he had a model, which we often call now the objectives-based model, for aligning the goals or objectives--what you're trying to achieve--with curriculum content and assessment. We find the standards, I think, given more prominence in our current model. In a sense, what Smith and O'Day have suggested is that it be applied to a much larger framework, maybe to a whole state, or possibly-conceivably-to a country; whereas Tyler thought in terms of a district, or even a classroom, or a school.

Now this amounts more to a specification of attributes that the reform should have than to a map for achieving reform itself. And this is necessarily the case. That is, in a sense the specification, the blue print, of what systemic reform should be is incomplete. And I don't see how it really could be any other way. You

cannot say exactly, here's what New York should do, or here's what Louisiana should do. Given this constraint, systemic reform can, and does, take many different forms. That's the specification of reform.

Then you turn to the real components. What I call the real components of the reform are departments of education, school districts, teachers, politicians, administrators—the actual actors and agents in the system itself. This is partly reflected in Barley's paper on Donald Schon's work on theory—the theory per se and the theory in use. So when you have the general specification of what systemic reform is, that's one thing; when it is put into practice or into use we have quite a different idea. It has to be implemented some how, so that you end up with all of these different manifestations. If you look at the SRI International's model of systemic reform—of what they finally had to address across states—they came up with a model that would put all the states on the same map. This model mostly consists of the actual program components, or the real agents. You can see that if you take out systemic reform activities on the one side, you could put almost anything in its place and you would still have the same actors, teachers, politicians, and departments of education. You could put your own activities in there, in a sense. Your own criteria. To what degree have you influenced teachers? It is very comprehensive. They needed a model that in some way would put all of this together. So, my question: How do we account for this particular state of things, having to look across states.

In these twelve papers, systemic reform sometimes refers to the original, theoretical specifications, and sometimes it refers to the real educational systems themselves. And often—as we have heard in the papers

presented-it's a mixture of the two: in some cases, of the design specifications, and, sometimes, the real components. So those get blended and mixed together. Hence, we have many different definitions and manifestations of systemic reform. Now this creates problems for the evaluators, for the 'evaluations, as it did for the original SRI evaluation, problems they had to work on for some time. Some evaluators use as criteria for the evaluations, for example, the more theoretical generic attributes of systemic reform-its systemicity, or systeminess, I think somebody said-or what Norman Webb calls alignment. The alignment of these things. This is one of the theoretical attributes that systemic reform has. So some of the papers lean heavily on that. The Webb paper is a good example of this emphasis.

But the criteria for evaluation that we have actually heard discussed, and those that have been proposed, come from many other sources. I think there are at least as many as nine sources, and maybe more, for deriving criteria for evaluations. First, you can look at the program itself, or the policy. Look at systemic reform, as it is specified. You can also consult the evaluation models, which some people have done. There are many of those. This is reflected mostly in the issue of what role the evaluator should play. Should the evaluator be a judge, or should the evaluator be a friend, all of those kinds of considerations. Those are reflections of the evaluation approaches themselves. You can also ask the potential audiences for the evaluation, including the clients, what should be the focus in the evaluation and what they want out of the evaluation. You can ask the other stakeholders, the people implementing the program, the teachers, what they expect and want out of the evaluation. You can determine the needs of the recipients. I mean, what is it (I would presume the students here are the recipients), what is it these students need?

You can also consider institutional points of view, when you are deriving criteria. By that I mean, you can look at it from a political angle, which has come up quite often. You can look at it from an economic angle. The

value-added approach is essentially an economic model. The idea of setting up an evaluation in such a way-even though we are talking about test scores in the end-we are still setting it up in the same way that economists use to look at value-added to financial components. Or, you could consider the context of the program and the context of the evaluation. Where does this program fit into the whole picture here? Or, where does the evaluation fit in? And who is involved in it? Finally, you could consider social theory, such as social justice or theories of gender. This has come up only peripherally, I think, in most of these twelve papers. I'm a bit surprised we didn't hear more of it. Concern about minorities, or concern about women, or other kinds of groups. We have not heard much of that here. I am not saying that people are not interested, but it has not emerged as an issue in most of these papers.

Now, if you have all of these possibilities for deriving criteria for the evaluation, what do you do? Well, different people have chosen to emphasize one or the other. Evaluators have to consider, when they are setting up an evaluation or proposing an evaluation, the audiences for the evaluation, and the purposes for the evaluation as well. That is, maybe we want to justify the program to Congress. And maybe the value-added analysis offers some steps in that direction because Congress, we know, is very interested in this kind of information. Or, maybe we want to look at whether the theory of systemic reform actually works or not. So, we end up with the kind of suggestions that Clune has made, which are really looking at whether the theory is functioning as the theory is supposed to function; this quantitative reduction, or that qualitative data. The academics tend to be mostly interested in theoretical aspects of it. Or, you could look at evaluations for local program improvements. Maybe the program logic? Are these people quite sure of what they're doing? Do they, can they, specify what they are doing? So you go through the program logic procedure that's been suggested. Or, maybe you're looking at justifying and improving local activities by feeding data back to the city or state, as in

Detroit, in Puerto Rico, or in Ohio. Thus, once they have chosen particular approaches to take, they need to determine what kind of information is valuable. And they have implemented those programs apparently with some consummate skill and results. So we find that each of these evaluation papers offers a reasonable response, more or less, to the specific evaluation challenges, contexts, and problems identified. Everyone, when they got to this point of saying what the evaluation is, they specified what systemic reform is in their view, and they also indicated what the evaluation should consist of. What I am saying, however, is that they have to fill in a lot of the blanks before they reach this point.

Mark St. John did one of the most remarkable turn-arounds in a sense; he redefined the context of the entire program, and the entire context of evaluation, and concluded by saying that what we can expect from evaluation is not justification, but understanding. The only thing we can really expect systemic reform to do is to promote capacity within the system as a whole, not to produce some kind of determinant outcomes, such as test scores. So when Mark filled in his blanks, he did a total kind of redefinition of things, compared to what a lot of other people have done. Thus, we find that, in a sense, there's no single best evaluation design that serves all purposes for systemic reform, given the nature of the objective under review, the many different contexts in which the reform is implemented, and the many different purposes that evaluation might serve. Each of the papers presents a different approach to the problem. So what can the National Science Foundation (NSF) do about this?

Well, maybe NSF can help. As I listened to people in the groups, they tended to worry about what they should be doing with their evaluations. They were asking: Are we doing the right thing? What should we be doing? Maybe NSF could provide a little more clarification on priorities. There are several ways they could do this. They could act through their advisory groups. They could act through their agents, like NISE. Or, they could hire consultants. There are many things they could do. They could have a meeting like

this to decide what the top priorities are for the evaluation and try to arrive at that through joint discussion. But one of the things that NSF could do is to establish priorities for what these evaluations are supposed to do. Now I personally believe you really need several tiers of evaluations. There are a number of levels of evaluation. I think you need an evaluation, or several, which serve to report to the Congress and to legislators on what is going on. You need certain kinds of information for that. You also need information at the local level in order to improve the program itself. That's my own bias. I would say this about other national systems of evaluation, those in other countries, that you need this two-tier level that is not necessarily very tightly connected, but that may reflect different concerns at different levels about the same ongoing program. Another thing NSF could do for its agents would be to construct a topology of designs to serve various purposes. That is, it could take some of these ideas and say here's a topology which shows that, if you're concerned about this, you go this way. If you're concerned about that, you might go that way. Or, NSF could also develop some model evaluations that show people at a given level what they could do. Here are some ideas. The more specific it is, the better it tends to be, and the more people tend to learn from it. You like to see some concrete material, even though you may do the design, or may have done the design, differently yourself.

Now, finally, there's a huge amount of commentary in the feedback from the small groups about student achievement and a fear, a concern, that student achievement will be used as the ultimate determinant of whether this works or not. People are not very happy with that idea, for the most part. Now, I think collecting student data is pretty much inevitable. I mean, at least to justify this to Congress and to legislatures and other decision makers. I think there are different ways to approach this that are not as harmful as what might at first appear. There are several things that could be done. Among them, you could have small experiments under ideal circumstances, like drug tests, to

see whether this alignment of curriculum and assessment really results in better test scores. You could make this kind of pure effort. We know this is not how the program is going to work out in the country; but, on the other hand, we **know** that people don't take their blood pressure medication, either. So you could attempt these small trials.

We could also try the kind of value-added approach I referred to earlier. There are problems with all of these approaches, of course, lots of difficulties with the **reliability**-of-gain scores in a statistical sense. So there are difficulties to be worked through, although I think that offers interesting challenges to the value-added approach. There are also ways of backing off the test scores just a bit, when you look at different forms of impact. For example, I got involved in trying to evaluate the impact of the CRESST (Center for Research on Evaluation, Standards, and Student Testing) Research and Development Center, which is a Department of Education R&D Center. CRESST is trying to promote alternative assessment, which some of you are already using in your designs in one form or another. What I did, rather than say the CRESST materials at this stage of development can't really be shown to improve achievement, I suggested starting with what

they are doing. What you can look at is what it would take for this alternative assessment to be developed and put into place. And so identify in the evaluation the gatekeepers, as it were, to getting an alternative assessment into place like the research community and the measurement community itself. I looked to see how they responded to alternative assessment, which is actually pretty positive. I also surveyed the state and local test directors to see how they responded to authentic assessment. They tended to be positive, with a major proviso: They did not want to give up standardized achievement testing. They wanted to do both. Which then presents the problem of cost. If you're going to do both, it will be very expensive. As we already heard from Ohio, they have abandoned some of their authentic assessment procedures because of the expense of those alone. This is difficult. It's a difficult problem. But it is possible to back off to show impact, if you want to talk **impact a bit-to show impact on different groups without necessarily converting this into achievement test scores at the end.** Now, ultimately, on the political level, we will have to provide some kind of evidence down the line. However this is formulated, it has to be done carefully to show achievement impact.

Closing Observations

Marshall Smith
Deputy Secretary
U. S. Department of Education

I come here with a distinct disadvantage, having not heard the papers you have presented during the last two days, which are steeped in education problems and issues that I am sure have been addressed, and addressed well in this Forum. I want to try to do something a little different obviously from Cora and Ernie, who were summarizing your discussions and your thinking, bringing their own insight to it. What I will try to do is talk some about the current environment for the evaluation of systemic reform. And then actually take this opportunity of setting out, in a series of steps, what I think might be done to evaluate systemic reform. In so doing, I would also like to discuss some ways that we typically do not think about in carrying out evaluations. My orientation here is less toward the National Science Foundation's Systemic Initiatives than it is toward the kind of effort, of course, that the Department of Education is undertaking and that Jennifer O'Day and I wrote about. This approach is based upon a model that is applicable across the different subject matter areas—i.e., it does not focus just on mathematics and science—and on a reform strategy initiated typically at the state level, not in a rural district, for example, or even in a big city, although it could be generalized to that, I think. As is the case in some of the state systemic initiatives for the National Science Foundation, I am going to be addressing the overall reach of systemic reform, rather than concentrating on one or two aspects of it. I know for, example, that in California, at least, the early state systemic reform focused on professional development per se—as a project that could be studied just for professional development. However, I want to talk particularly about trying to evaluate across-the-board in systemic reform.

First, let me sketch for you what I think systemic reform is. Clearly, states differ on

systemic reform, or standards-based reform, whichever you might call it. But I want to argue that there's a general framework here and that general framework has some external drivers behind it—Goals 2000, to some extent, Title I and the Elementary and Secondary Education Act, to a greater extent. This is not to suggest that the nation, and the states, would not have moved in this direction were it not for Goals 2000, or Title I; but it is to say that under a law such as Title I, or a directive like Goals 2000, the states were expected, and agreed, to follow a sequence, a set of practices, both Ernie House and Cora Marrett have mentioned here. I know that you have all thought about these practices.

Let me review quickly. A first step is that states develop content and performance standards, for a variety of different academic areas. And, that those content and performance standards are focused on all children, *all children*. I want to stress that as much as I conceivably can. Jennifer and I wrote a second paper, published in 1993, I believe, which focused specifically on equality in systemic reform, where we argued, I think fairly cogently, to the effect that unless we do have some common set of standards that apply across an entire state, we are going to continue to have the kinds of disparate curricula—one curriculum for kids in the inner cities and one curriculum for the kids in the suburbs—that we have had for too long in our nation. The kind of emphasis that we are talking about is an equalizing emphasis, or could be an equalizing emphasis. So one of the major outcomes that I would like to see come out of this effort is a closing of this gap: a closing of the gap in achievement levels between blacks and whites, between Hispanics and whites. Also, I want to see an increase in all scores. Without that, I would consider this reform to be a failure. One component is state-level content and

performance standards and aligned assessments. Aligned means that the assessments and the content and performance standards are congruent. It means that what the assessments measure is based upon that content and those performance criteria that are outlined in the standards. We do not yet know a great deal about alignment. That is a fundamental research question that we have to focus on. We have some ideas. We can sketch out matrices and a variety of other things, but we are not very good at it, and we do not know enough about it. On the other hand, I think judgments can be made by experts, looking at these materials, about a lack of alignment or greater or lesser alignment, but it's not easy.

Second, resources, not just the assessments, must be aligned. That is, resources aligned to support the students in their efforts to meet the standards, teachers prepared to teach to the standards, both through pre-service and in-service training; local curriculum and instruction aligned; and other resources needed to support the effort. Local, district, and state flexibility is essential in adopting strategies for helping all children achieve the standards. We must find some ways to give local schools and districts the kind of flexibility and resources during their reform effort that will enable them to adapt their strategies to the particular needs of their students. And, finally, student and school performance accountability is to be based on aligned assessments.

Not only is there a generally agreed-on framework, but there is also a general time sequence in implementing the components. States have started from different places. The politics has varied in states. And, states have progressed somewhat differently. But, by and large, the sequence holds up across the states. The content standards generally come first. There are 48 states now with content standards. Performance standards and assessments seem to be coming at one and the same time. Approximately 25 to 30 states have performance standards. Roughly 20 states have informed the Council of Chief State School Officers that they have aligned assessments; others are nearly there—are

really ready, but just not online yet. So we are moving through those stages. There is a lot of talk in states about aligning their teacher development, professional development, and some of their activities in their teacher training institutions with the content and performance standards. That may be the hardest nut to crack.

But you can see that resources are beginning to line up, more or less. And it does vary by state. Accountability is taking on new importance. It's interesting: If you have been reading the *Washington Post* over the last three weeks, you have read a lot about Virginia and its new SOLs (Standards of Learning), and its accountability mechanisms. If you have been reading *The New York Times*, you are reading a lot about the grade 4 assessments in New York State. And if you have been reading the Dallas newspapers and the Houston newspapers, you have been reading about Texas accountability. There is a lot of interest in student accountability and in the issue of accountability with respect to failing schools. There is interest, also, in a variety of other things. So, you now have a sequence. You have not just a structure, but also a sequence.

Third, you have different kinds of evaluators and assessors out there. We have *Education Week* evaluating. We have the American Federation of Teachers, *Achieve*, and the National Science Foundation. The Consortium for Policy Research in Education (CPRE) is out evaluating. There are evaluations by the Goals 2000 Panel, the RAND report on Texas and North Carolina, which I commend to you if you have not seen it. The evaluations differ in focus, in rigor and content, but they are all evaluations of these particular reforms.

Fourth, this refers more to the environment right now. In the last two years, we find ourselves in a new era. Five years ago most of us did not use the World Wide Web very much. Now we use it all of the time. Communication is ubiquitous; information and data can be transmitted instantaneously. People are not willing to settle for the old time frames for evaluation; they want new time frames. They want information and they want

data on their time frame, not on the time frame of the evaluators. This represents a major difference in the way evaluations are now working. The whole sense of science and information in public policy has speeded up in a way that we could not have imagined possible a few years ago. You can now get instantaneous data and put it on the Web. You can go to the Web—quite extraordinary—you can go to the Web and get thousands and thousands of school report cards now, which contain school achievement data on them. In three or four years, you are going to be able to go back and look at three or four years of data on those report cards. And this is just the beginning. This is the last point. We are at the beginning of an era when we will have an incredible mass of rich information. We typically call distributed data bases what we can begin to link up now via the Web. These are the source of incredibly valuable information for us. This is not just information that is unmassaged. It is also evaluation reports from all of you who can go on the Web, from your graduate students who can get on the Web, and from graduate students from universities that you have never heard of who can get on the Web. They can all be studying this same set of phenomena. There are huge potential masses of data out there that are at your beck and call.

So what does all of this suggest about how systemic reform might be evaluated? Ernie House's point is probably the most important one: You have to figure out what you want out of the evaluation. What is the purpose of this whole thing? If you want summative data from across the nation for any given period of time, six months, a year, two years, and you want to look at absolute performance levels and gains vis-a-vis closing the gap, you better settle for NAEP (National Assessment of Educational Progress) data, and ignore the alignment problem. If you have NAEP data and *Education Week*, and some aspirin, you can probably make some pretty cogent arguments to Congressmen and legislators for education change. This is going to be done no matter what happens. People are going to do this kind of analysis. So I do not think you all need to worry about it too much.

We ought to be sure about what we technically call the unit of analysis. Because in my model, at least, and in the way I think about it—and we are thinking about it—the reforms are state based. You can think of districts and schools as being roughly in a hierarchy. Imagine a hierarchical model, and you can begin to aggregate upward within that model, if you wanted to start at the bottom. Or, you could even begin to aggregate across some of the states. And here, there are a number of different components. Clearly, since we have a structure and a time sequence, we can begin to determine whether that structure is in place, and whether it is getting into place in roughly the time frame set by the state, or that set by the Title I regulations, or whatever the parameters that apply. So we begin with a pretty explicit set of characteristics, fundamental building blocks, that are being put in place. Are there content and performance standards? What other qualities? For what purposes? Evaluation slides into research at some point.

Again, we do not really have a theory about content standards and about how to define the breadth and depth of content standards, or of performance standards, for that matter. We have not yet defined their relationship. We do not have an understanding of the relationship between content standards and assessments and the nature of alignment. Nonetheless, we have lots of folks out there who are evaluating the quality of content standards. So, that could be looked at. That is pretty straightforward. Are there aligned assessments? Again, can you evaluate the quality of those assessments as you begin to move through the sequence? Are you beginning to see a distribution of alignment resources? Are the resources being delivered in a fair manner across the state? In some sort of equitable manner? Are decisions in higher education being made with consideration of the reforms and the student standards in mind, or with no consideration of these factors? Are schools of education programs tracking the student standards? Are accountability systems being developed? Are they fair? Are local schools and districts being provided the support and flexibility they need in order to

respond to local needs? All of these items can be tracked over time. They should be tracked over time. They should be tracked for the details in the data. They should also be tracked simply to provide feedback information on a regular basis.

At a deeper level, we can ask the question: Are the reforms seeping into the consciousness of the key actors—of the parents, of the students, of the teachers, of the principals and so on? Do teachers know about the reforms and standards? Do the parents? Do the people in central office? Are networks of teachers developing? You also can get into something slightly more subtle, networks of teachers beginning to develop over the delivery of instruction that begins to meet the standards. Are there conversations in the teacher's lounge? Are they talking about standards? Are they thinking about them? Are people tapping into Web pages that deal with the particular standards from a particular state? Are there self-generated sets of activities that teachers and principals and others engage in to think about standards? Can we detect changes in the character and content of teaching in the classrooms? The questions begin to get deeper and deeper. These are fairly obvious deepnesses. You can move from the concrete, and into things that are a little more subtle. What sort of strategies, other strategies, are being used to move all students to higher standards, beyond what's going on in the classrooms, summer schools, after schools, and so on? What are the unintended consequences of all of this? And, if one wanted to really move to a point of much deeper understanding—something that Cora implied in one of her **comments**—are these systems beginning to self-organize? Are they operating as organic systems, in the way the brain operates, or that other complex systems operate? Are the linkages and the common understandings present in such a way that the system evolves naturally, without prodding by the government or others? These questions are amenable, many of them, to on-time, real-time studies, amenable to providing clear information for feedback—feedback to the teachers, to the principals, to the people making the decisions, so they can, in fact,

make more valid judgments, policy judgments at their level, about what to do with resources, and about how to respond. Is the gap closing in the district? Are certain teachers in one set of schools, the more advantaged schools, beginning to change their teaching policies, while the teachers in the other schools are not? And is that operating to the disadvantage of certain kids? Are certain schools getting the kinds of textbooks needed earlier than others, when the textbooks are necessary for students to really try to understand the materials? All of these issues relate to the nitty gritty, the fundamentals that are of utmost **importance**—that are absolutely key if these kinds of reforms are ever to have a chance to work

Throughout this process, you can measure achievement levels with unaligned tests. Achievement levels can be monitored, as well as other student outcomes, such as graduation rates, college admission, gap-closing efforts, and so on. As time goes on and aligned assessments come in, you could have a much better set of measures of achievement. You can establish initial benchmarks and begin to consider valued-added approach to your testing efforts, and reasonable accountability systems can become much more available. Again, it is absolutely critical that we begin to have fast turn-around on these things. It is something that you can help with in important ways. As you all know, most achievement tests are given in March, April, or May. Typically, test results are not returned until the next fall. Sometimes that means going from sixth grade to seventh grade, a totally different school; the kids scatter across districts. An evaluator should be on hand who can say, "Look, this isn't reasonable." Or, evaluators should make an observation, when they enter a district or a state, about student achievement. It does not help anyone to have these test scores available four months later. If they come in a month later, or so, the teachers can sit down and work with the students. They can help to guide them, so they are ready for the next grade. Finally, during the summer, the school can work on making changes on its own program, make efforts to improve the quality of its programs.

Now, as you know, what I have just sketched operates in something of a **never-never** land. That may be the academic in me. **But there** is a structure out there. I believe the world is a lot more messy, however, than that. I use a slightly different metaphor. Not Cora's boat rocking through the storm. Road maps, developed on the way, can give us guidance while we move along some of the pathways, byways, and detours that we are going to encounter. It is that kind of road map, though, that kind of vision, that connects both understanding the sequence-the structure and the sequence-and the need to feed back information in to the system. That is absolutely critical.

Let me point out, however, three or four other factors that I think are changing, and helping to change what we can do. I have already mentioned the World Wide Web. One is that evaluations-I am sure you all have dealt with this, and have talked about this over the last couple of days-are tremendously advantaged if they use what I call benchmark variables for relationships. By that, I mean variables that people have developed and used over time. That is, the same question on a questionnaire, the same way of observing teachers, or whatever. Where there are good data about the distribution of those particular variables, depending upon the population that you are studying or the particular sub-sample, that would be a benchmark variable. A benchmark relationship is one in which there would be good information about the correlation, about some sort of regression coefficient, some measure of a relationship between two variables, or multiple variables, so that you can pick them up out of other studies and put them into your study, and use them as a way of trying to understand the new data that you have obtained. Because if you have established some benchmark relationships, some benchmark variables in relationships, you can then see whether they differ in your own study. And if they do differ in your study, you have to sit back and worry about this a little.

What happens if they differ in your study-I mean, what are your hypotheses? Well, one of my hypotheses is that your data

are lousy. Right? They are biased somehow. They do not work. They do not give you the same relationship. Another is that you have developed some sort of sample-by-sample, or population-by-population interaction. Now, that's interesting and you have to explore it. You probably have to pick up another body of data to look at it, or try some other benchmark relationships. Another possibility is that the relationship itself has changed. The relationship between schooling, for example, and future income has changed over the last thirty or forty years, which explains some of the differences in economic studies that had looked at the productivity of schooling-a difference that has increased over time. But the idea here is to use former studies to build validity into your studies. And it is absolutely critical for you to do it. It is done far too seldom in our research.

A second possibility is to use distributive data bases-to begin to build them and use them. They are going to be increasingly available across states and districts. There ought to be money from the National Science Foundation and from our department and other sources to make them available to people all over the country. The third strategy I recommend here is to synthesize, synthesize, synthesize. To look at other people's work, to try to match it against your work. To think, to open your eyes to other kinds of research that relate to yours. How does the research on the reconstruction of industry, of businesses, and so on, relate to systemic reform? It relates, I think, very directly. How does the work on complexity theory and chaos theory, which Cora and I were just mentioning indirectly, how does complexity theory relate to systemic reform in education? Extraordinarily enough, complexity theory-something that people call too systematized-the work on complexity theory and chaos theory, which comes out of the hard sciences, actually relates very directly to these reforms. As I mentioned, if we had a set of conditions under which we could create reform that **then** enabled organic growth in educational activity that involved all students achieving, we would have something truly wonderful.

Finally, the last point is that for most purposes-and here you will see my bias: you want to get on the clock of the people who are carrying out the reforms. Do the evaluations

on their clock (timelines), so that the evaluations can be useful to them rather than on your clock, or the academic clock, or even the funder's clock.

1999 Forum Observations-Handout

Marshall Smith
Acting Deputy Secretary
U. S. Department of Education

1. States differ on a variety of aspects of systemic/standards-based reform BUT there is a generally agreed-upon framework-and some external drivers—GOALS2000/Title I.
 - a. State-level content and performance standards and aligned assessments-focused on ALL students.
 - b. Resources aligned to support students learning to the **standards**—teachers prepared to teach to the standards, local curriculum and instruction aligned, other resources in support.
 - c. The flexibility of local schools and districts to adopt strategies for helping all students to achieve to the standards.
 - d. Student and school performance accountability based on assessments.
2. Moreover, there is a general sequence over time in implementing these components: States have started from different places and progressed somewhat differently, but the sequence generally has been standards first, then assessment and accountability, alignment happening throughout the time period. And in Title I, there are actual time lines for these efforts. **Forty-eight** states have content standards, some 25 or so performance standards.
 - a. States moving toward aligned assessments-by CCSSO count, up to 20.
 - b. Accountability taking on new importance-VA, NY, failing schools, TX, report cards. . .
3. Lots of different kinds of evaluators and researchers. ED Week, AFT, ACHIEVE, NSF, CPRE, the Evaluation Office in the Department, Goals Panel, Rand report, and so on. These evaluations differ in focus, rigor, and intent. Some focus on content and nature of the standards-some report, some evaluate.
4. We are in the eras of PERFORMANCE evaluations and reports (**GPPE**)—People want results and indicators of results that provide useful information.
5. We are also in the era of the Internet and World Wide Web. Data can be instantaneously transmitted; there is ubiquitous communication, people are not willing to settle for the old time frames for evaluations-they want quick, useful information. They want data on their time frame, not on the evaluator's time frame.
6. There are large bodies of information beginning to be available in rich distributed data bases throughout the nation. There are massive bodies of information about states and schools, including achievement and other information. Over time, with effort, these data bases could be organized to provide powerful evaluative information.

So what does this all suggest for how Systemic Reform might be evaluated?

 1. First of all, you need to know what you want out of your evaluation.
 2. Do you want summative data across the nation, assessing the overall effects of

the reforms? After six months, one year, five years, two generations? Here the data would look at both absolute performance levels, gains, and closing the gap. Good luck, no matter what the time frame. You better settle for NAEP (ignore the alignment problem) and ED Week and some aspirin.

3. If you want to be a little more thoughtful, focus on states as your unit of attention-after all, the reforms are generally state-based. You can think of districts and schools as being down the ladder in a hierarchical model. And you can aggregate up to get a sample that might generalize to the nation.
4. Since the reforms have a structure and a time sequence, even though there is state-to-state variation within the general framework I set out above, we can imagine a sequence of evaluative questions on the input and process side-questions that track the same time lines:
 - a. Are there content and performance standards? What are their qualities? For what purposes? Evaluation slides into research at some point: Do we have a theory about the depth and breadth of the content in the standards? How often should standards be specified? (By grade level or at natural breaks in schooling?)
 - b. Are there aligned assessments? Are they good assessments-For what purposes? What kinds of feedback are available from the assessments.
 - c. Is there the beginning of alignment of resources? Are the resources being fairly delivered? Are decisions in higher education being made without consideration of the reforms? Do School of Education programs track student standards?
 - d. Are accountability systems being developed? Are they fair? Are the incentives right?

Is the motivation to study the content in the domains, or the specific content on the tests?

- e. Are local schools and districts being provided the support and flexibility they need to respond to local needs?
 - f. All of these items can be tracked over time: Are deadlines and due dates being met? Is there qualitative improvement?
5. At a deeper level, are the reforms seeping into the consciousness of the key people and are deeper practices changing?
 - a. Do teachers know about the reforms and the standards? Do parents? Do the people in the central offices? Do the folks in higher education? Are they using the language of the reforms?
 - b. Are networks of teachers developing that focus on providing the kind of instruction that helps bring all students to reach the standards?
 - c. Can we detect changes in the character and content of teaching in the classroom-at different levels and in different situations?
 - d. What sorts of other strategies are being used to move all students to higher standards?
 - e. What are the unintended consequences of all of this?
 6. Questions amenable to on-time, real-time studies that provide high quality feedback information to help improve the evaluations-formative evaluation data that, for the most part can be quickly gathered, and made available over the Web and fed back into the system.
 7. Throughout this process, achievement levels can be monitored, as well as other student outcome indicators, such as graduation and college-going rates, gap-closing, etc.

8. As time goes on and aligned assessments come on board, benchmarks would be established and achievement growth tracked over time-and presumably, the accountability systems would kick in for real.
9. Now, the world is a lot more messy than this; but I believe that **roadmaps** can be developed along these lines that make it possible to explore alternative routes and byways without losing your way.
10. Above all, as we design evaluations, we need to think hard about what advantages we have that we have not had or not used enough in the past.
11. One advantage is the wealth of data and relationships that have been established in prior studies. A lot of this comes back to theory, or at least to **benchmarked variables and relationships**. A benchmarked variable is simply one for which you have prior consistent data about its value or values (e.g., a correlation or joint distribution or regression coefficient). Benchmarked variables and relationships such as those used in the status attainment studies in Wisconsin and Michigan serve as cornerstones for understanding new relationships and for testing new theory. With the same basic core data in multiple studies, you can begin to aggregate data and findings: Without such data, aggregation is very difficult. If your benchmarked variables and

relationships differ from the past, you have also learned something. Three possibilities occur:

- Your data and values are lousy and biased somehow.
 - You have discovered a **sample-by-sample** interaction, or even a **population-by-population** interaction.
 - The relationship has changed over time, like the relationship between test scores and future income changed from the 60s to the 90s.
12. A second advantage is the distributed databases to do all of the above, if possible. Keep improving the distributed databases-they are a big part of your future.
 13. A third strategy is synthesize, synthesize, synthesize. You should have the most eclectic, creative, and articulate synthesizers in the nation working for you. Draw on data beyond the data you collect, if at all possible. It is cheaper, faster; there is more of it.
 14. Finally, get on the clock of the people who implement the reforms so that your data are useful. Provide, in readable form, as much data as you can on the Internet so that it can be quickly integrated.

Appendix A Participant List

Joan Abdallah
AAAS
EHR
1200 New York Avenue, NW
Washington, DC 20005
Work: (202) 326-6673
Fax: (202) 371-9849
jabdalla@aaas.org

Nancy Adelman
Policy Studies Associates, Inc.
1718 Connecticut Avenue, NW
Suite 400
Washington, DC 20009

Andrew Ahlgren
AAAS
Project 2061
1333 H Street, NW
Washington, DC, 20005
Work: 2023266624
Fax: 2028425196
aahlgren@aaas.org

Ethan Allen
University of Washington
Molecular Biotechnology
Box 357730
Seattle, WA 98195
Work: (206) 221-4692
Fax: (206) 685-7301
ethana@u.washington.edu

Bernice Anderson
NSF
4201 Wilson Boulevard
Room 855
Arlington, VA 22230
Work: (703) 306-1650
Fax: (703) 306-0434
banderso@nsf.gov

Karen Anderson
American Psychological
Association
Center for Psychology in
Schools & Education
750 First Street, NE
Washington, DC 20002
Work: (202) 336-5860
Fax: (202) 336-6130
kanderson@apa.org

Susan Anderson
NCISLA
1025 W. Johnson Street,
Room 576A
Madison, WI 53706
Work: (608) 265-5630
Fax: (608) 263-3406
anderso@facstaff.wisc.edu

Gretchen Andreasen
University of California
Education Department
Merrill Trailers, UC
Santa Cruz, CA, 95064
Work: (831) 459-5770
Fax: (831) 459-5848
hampt@es.ucsc.edu

David Andrews
California State University,
Fresno
Biology
2555 E. San Ramon Avenue
Fresno, CA 93740-8034
Work: (209) 278-2412
Fax: (209) 278-3963
davidan@csufresno.edu

Martin Apple
Council of Scientific Society
Presidents
1155 16th Street, NW
Washington, DC 20036
Work: (202) 872-4452
Fax: (202) 872-4079
cssp@acs.org

Richard Audet
Roger Williams University
Education
One Old Ferry Road
Bristol, RI 02809
Work: (401) 254-3357
Fax: (401) 254-3286
rha@alpha.rwu.edu

J. Martin Ball
Indiana Dept of Ed
Indiana Dept of Ed, Rm. 229
State House
Indianapolis, IN 46204
Work: (317) 232-9112
Fax: (317) 233-9121
maball@doe.state.in.us

F. Robert Barak
Iowa State University
Board of Regents
IKE 206-13-10
E. 12th & Grand
Des Moines, IA 503 19
Work: (515) 281-3939
Fax: (515) 281-6420
rbarak@iastate.edu

Zoe Barley
Co-Director
Western Michigan University
Science & Math Program
Improv.
3225 Wood Hall
Kalamazoo, MI 49008
Work: (616) 387-3791
Fax: (616) 387-3770
barley@wmich.edu

Lehman Barnes
University of North Florida
Higher Ed Consortium for
Mathematics & Science
2335-201 Costa Verde Blvd.
Jacksonville Beach, FL, 32250
Work: (904) 620-1074
Fax: (904) 620-1025
lbarnes@unf.edu

Marianne Barnes
University of North Florida
Curriculum & Instruction
2335-201 Costa Verde
Boulevard
Jacksonville Beach, FL 32250
Work: (904) 620-2578
Fax: (904) 620-1025
mbarnes@unf.edu

Came Bamett
WestEd
Mathematics Case Methods
Project
500 12th Street, #340
Oakland, CA 94607
Work: (510) 587-7329
Fax: (510) 587-7373
cbamett@wested.org

Lynn Barnett
American Association of
Community Colleges
Academic, Science, &
Community Development
One Dupont Circle, #410
Washington, DC 20036
Work: (202) 728-0200
Fax: (202) 833-2467
lbarnett@aacc.nche.edu

Phyllis Bamhart
West Virginia Dept of Education
Office of Instructional Services
1900 Kanawha Blvd., East,
Building 6, Room 330
Charleston, WV 25305
Work: (304) 558-7805
Fax: (304) 558-0459
pbarnhar@access.k12.wv.us

Constance Barsky
The Ohio State University
Department of Physics
174 W. 18th Avenue
4138 Smith Lab
Columbus, OH, 43210
Work: (614) 292-3323
Fax: (614) 292-3221
barsky.1@osu.edu

Dennis Bartels
The Exploratorium
Center for Teaching & Learning
3601 Lyon Street
San Francisco, CA 94123
Work: (415) 353-0496
Fax: (415) 561-0307
dbartels@exploratorium.edu

James Barufaldi
University of Texas-Austin
Science Education Center
1912 Speedway SZB356
Austin, TX, 78712
Work: (512) 471-7354
Fax: (512) 471-9244
jamesb@mail.utexas.edu

Michelle Batchelder
Austin Independent School
District
Office of Program Evaluation
1111 West 6th Street
Austin, TX 78703
Work: (512) 414-3565
Fax: (512) 414-1707
mbatchel@admin.austin.isd.tenet. edu

David Bauman
Capital Area Institute for
Mathematics & Science
55 Miller Street
P.O. Box 489
Summerdale, PA 17093
Work: (717) 732-8427
Fax: (717) 732-8414
dbauman@cauii.k12.pa.us

Sharon Beckstrom
RMC Research Corp.
Regional III Comprehensive
Center
1815 N. Fort Myer Drive, #800
Arlington, VA, 22202
Work: (703) 558-4800
Fax: (703) 558-4823
beckstrom@rmcarl.com

James Bernhard
University of Massachusetts
Donahue Institute
100 Venture Way, 3rd Floor
Hadley, MA 01035
Work: (413) 587-2404
Fax: (413) 587-2410
jbernhard@external.umass.edu

Katie Bickel
Horizon Research, Inc.
111 Cloister Court, Suite 220
Chapel Hill, NC, 27514
Work: (919) 489-1725
Fax: (919) 493-7589
kbickel@horizon-research.com

Larry Bilbrough
NASA Headquarters
Education
Code FE
Washington, DC 20546
Work: (202) 358-1439
Fax: (202) 358-3048
larry.bilbrough@hq.nasa.gov

Eve Bither
U.S. Department of Education
OERI
80 F Street, #100
Washington, DC 20208-7564
Work: (202) 208-0692
Fax: (202) 219-1528
eve-bither@ed.gov

Suzanne Blanc
Research for Action
3701 Chestnut Street
Philadelphia, PA, 19104
Work: (215) 823-2500
Fax: (215) 823-2510
sukeyblanc@aol.com

Rolf Blank
ccsso
Evaluation Indicators
1 Mass Avenue, NW
Washington, DC 20001
Work: (202) 336-7044
Fax: (202) 789-1792
rolfb@ccsso.org

Elizabeth Boehm
San Antonio USI
110 Tuleta
San Antonio, TX, 78212
Work: (210) 734-0016
Fax: (210) 734-7890
eboehm@tenet.edu

Becky Bogert
Northern Arizona University
Science & Math Learning Center
P.O. Box 5697
Flagstaff, AZ 86011
Work: (520) 523-7160
Fax: (520) 523-7953

Denise Borders
The Mackenzie Group
1100 17th Street, NW, #1100
Washington, DC 20036
Work: (202) 466-1111
Fax: (202) 466-3363
bordersd@mckgrp.com

Patricia Bourexis
The Study Group
209 Sir Walter Raleigh Drive
Kill Devil Hills, NC 27948
Work: (252) 441-2788
Fax: (252) 441-9663
study_group@aol.com

Andrea Bowden
Baltimore City Public School
System
Office of Science, Mathematics
& Health Ed
200 East North Avenue
Baltimore, MD 21202
Work: (410) 396-8585
Fax: (410) 396-8063

Robert Brazzle
North Central Regional Ed. Lab
Midwest Math & Science
Consortium
1900 Spring Road, # 300
Brook, IL 60523
Work: (630) 218-1261
Fax: (630) 571-4716
bbrazzle@ncrel.org

Edward Britton
NCISE
1726 M Street, NW, #704
Washington, DC, 20036
Work: (202) 467-0652
Fax: (202) 467-0659
britton@ncise.org

Dennis Brown
Michigan State University
4129 Schoads Drive
Okemos, MI 48864
browndepiplot.msu.edu

Sharon Brown
Cleveland Education Fund
1422 Euclid Avenue, Suite 1550
Cleveland, OH 44115-2001
Work: (216) 566-1136
Fax: (216) 566-1230
sabrown@cleveland.ed.fund.org

Bonnie Brunkhorst
California State University-San
Bernardino
Science, Mathematics, &
Technology Education
Department
6288 Alegre Ct
Riverside, CA 92506
Work: (909) 780-6545
Fax: (909) 780-3640
bbrunkho@csusb.edu

Herbert Brunkhorst
California State University-San
Bernardino
Science, Mathematics, &
Technology Education
Department
5500 University Parkway
San Bernardino, CA 92407
Work: (909) 880-5613
Fax: (909) 880-5988
hkbrunkh@csusb.edu

Pamela Buckley
Eisenhower Regional
Consortium for Mathematics
& Science Education
1700 N. Moore Street, # 1275
Arlington, VA 22209
Work: (800) 624-9120
Fax: (703) 276-0266
buckleyp@ael.org

Joseph Burke
Miami-Dade County Public
Schools
Division of USI Mathematics &
Science
1500 Biscayne Boulevard
326B
Miami, FL 33132
Work: (305) 995-2341
Fax: (305) 995-1916
jburke@dcps.k12.fl.us

Rodger Bybee
National Research Council
Center for Science, Mathematics
& Engineering, Education
2101 Constitution Avenue, NW
Washington, DC 20418
Work: (202) 334-2353
Fax: (202) 334-2210
rbybee@nas.edu

Linda Cain
Oak Ridge National Laboratory
Office of University & Science
Education
Oak Ridge National Laboratory
Building 5500A, MS 6356
PO Box 2008
Oak Ridge, TN, 37831-6356
Work: (423) 576-3886
Fax: (423) 241-6776
cainlc@ornl.gov

Barrett Caldwell
NISE
1025 W. Johnson Street
Room 753C
Madison, WI 53706
Work: (608) 263-9250
Fax: (608) 262-7428
caldwell@engr.wisc.edu

Dave Calhoun
Fresno Unified School District
Research, Evaluation &
Assessment
2306 Tulare
Fresno, CA, 93721
Work: (209) 441-3569
Fax: (209) 265-2913
docalho@fresno.k12.ca.us

Barbara Cambridge
American Association for
Higher Education
Teaching Initiatives
One Dupont Circle, #360
Washington, DC 20036
Work: (202) 2936440
Fax: (202) 293-0073
bcambridge@aahe.org

Patricia Campbell
University of Maryland, College
Park
Department of Curriculum &
Instruction
2226 Benjamin Building
College Park, MD, 20742
Work: (301) 405-3129
Fax: (301) 314-9055
pc2@umail.umd.edu

Jeanne Rose Century
Education Development Center
Center for Science Education
55 Chapel Street
Newton, MA 02158
Work: (617) 969-7100 x2414
Fax: (617) 630-8439
jeannec@edc.org

Karen Cerwin
CA K-12 Alliance
P.O. Box 5
Rim Forest, CA, 92378
Work: 9093379131
Fax: 9093379131
kcerwin@cams.edu

Juanita Clay Chambers
Detroit Public Schools
Office of Mathematics &
Science
5057 Woodward Avenue
Room 932
Detroit, MI 48202
Work: (313) 494-1082
Fax: (313) 302-2440
juanita_chambers@dpsnet.detroit.k12.mi.us

Jeffrey Char-vet
AAAS
Directorate for Education &
Human Resource Program
1200 New York Avenue, NW
Washington, DC, 20005
Work: (202) 326-6644
Fax: (202) 371-9849
jcharvat@aaas.org

Anne Chase
ABT Associates, Inc.
55 Wheeler Street
Cambridge, MA 02138
Work: (617) 349-2453
Fax: (617) 349-2665
anne_chase@abtassoc.com

Daryl Chubin
Senior Policy Associate
National Science Foundation
National Science Board
4201 Wilson Blvd. Rm. 1225
Arlington, VA 22230
Work: (703) 306-1650
Fax: (703) 306-0434
dchubin@nsf.gov

James Cibulka
University of Maryland
Education Policy, Planning &
Administration
2115 Benjamin Bldg.
College Park, MD 20742
Work: (301) 405-3589
Fax: (301) 405-3573
jc292@umail.umd.edu

Paul Cieslak
Milwaukee Public Schools
Research & Assessment
5225 W. Vliet Street
Milwaukee, WI 53108
Work: (414) 475-8257
Fax: (414) 475-8262
cieslapj@mail.milwaukee.k12.wi.us

Marianne Cinaglia
Rowan University
Secondary Education
201 Mullica Hill Road
Glassboro, NJ, 08028
Work: (609) 256-4500
Fax: (609) 256-49 18
cinaglia@rowan.edu

Susan Cloutier
Lesley College
PERG
7 Evergreen Avenue
Wellesley, MA 02482
Work: (781) 235-2611
cloutier@mediaone.net

William Clune
NISE
1025 W. Johnson Street
Room 753F
Madison, WI 53706
Work: (608) 263-4348
Fax: (608) 262-7428
clune@macc.wisc.edu

Susan Cohen
Lesley College
PERG
29 Everest Street
Cambridge, MA 02138
Work: (617) 349-8458
Fax: (617) 349-8668
suecohen@mail.lesley.edu

Kathy Comfort
WestEd
California Systemic Initiative
Collaborative
730 Harrison Street
San Francisco, CA, 94107
Work: (415) 565-3061
Fax: (415) 565-3012
kcomfort@wested.org

Deborah Cook
Rutger University-Bosch
Campus
NJ SSI
640 Bartholomew Rd.
Piscataway, NJ 08854
Work: (732) 445-2249
Fax: (732) 445-2848
cookd@dimacs.rutgers.edu

Francena Cummings
Eisenhower Consortium
Mathematics & Science
1203 Governors Square Boulevard
400
Tallahassee, FL 32301
Work: (850) 671-6033
Fax: (850) 671-6010
fde3530@garnet.acns.fsu.edu

Delores Dalton
Virginia Department of
Education
Secondary Instruction
P.O. Box 2120
Richmond, VA, 23218-2120
Work: (804) 371-0778
Fax: (804) 371-5466
dudalton@pen.k12.va.us

Marita Danek
COSMOS Corporation
3 Bethesda Metro Center, #950
Bethesda, MD 20814
Work: (301) 215-9100
Fax: (301) 215-6969

Deborah Dardis
Southeastern Louisiana University
Biological Science
SLU 10788
Hammond, LA 70402
Work: (504) 549-2163
Fax: (504) 549-3851
ddardis@selu.edu

Kerry Davidson
Louisiana Board of Regents
150 N. Third Street, # 129
Baton Rouge, LA, 70801-1389
Work: (225) 342-4253
Fax: (225) 388-0688
davidson@regents.state.la.us

Norma Dhila
University of Puerto Rico
Department of Psychology
Box 23334
San Juan, PR 00931-3334
Work: (787) 765-5170
Fax: (787) 756-77 17
n_davila@uprl.upr.chi.edu

Darnella Davis
COSMOS Corporation
3 Bethesda Metro Center, # 950
Bethesda, MD 20814
Work: (301) 215-9100
Fax: (301) 215-6969
ddavis@cosmoscorp.com

Reeny Davison
ASSET Inc.
290 Corliss Street
Center City Terminal
Pittsburgh, PA, 15220
Work: (412) 771-2121
Fax: (412) 771-5130
rdavison@lcn.net

Melanie Dean
UC, San Diego
Program for Teacher
Enhancement in Science
& Technology
9500 Gilman Drive
La Jolla, CA 92093-0176
Work: (619) 534-8587
Fax: (619) 534-7483
mdean@ucsd.edu

Robert Dean
University of California, San Diego
Program for Teacher Enhancement in Science & Technology
9500 Gilman Drive
UNEX/PTEST 0176cc
La Jolla, CA, 92093-0176
Work: (619) 534-8587
Fax: (619) 534-7483
rdean@ucsd.edu

Robert DeHaan
Emory University Medical School
ESEP Program
575 Rollins Way
Atlanta, GA, 30322
Work: (404) 727-3050
Fax: (404) 727-3051
bob@cellbio.emory.edu

Sharon Derry
UW-Madison
NISE
1025 W. Johnson Street
Madison, WI 53706
Work: (608) 263-3676
Fax: (607) 263-6448
sderry@macc.wise.edu

LaDonna Dickerson
NISE
1726 M Street, NW, #704
Washington, DC, 20036
Work: (202) 467-0652
Fax: (202) 467-0659
dickerso@ncise.org

Kathy DiRanna
CA K-12 Alliance
8565 Rhoads Circle
Fountain Valley, CA 92708
Work: (949) 824-7809
Fax: (949) 824-7621
kathy_diranna@carns.edu

Joan Donahue
National Alliance of State Science & Math Coalitions
11 Dupont Circle, NW
Suite 250
Washington, DC, 20036
Work: (202) 387-3600
Fax: (202) 387-4025
jdonahue@nassm.org

Sally Donlon
Louisiana, Board of Regents
University of Southwestern Louisiana
USL Box 43688
Lafayette, LA 70504-3588
Work: (318) 482-1018
Fax: (318) 482-2051
sod@usl.edu

Carolyn Domsife
University of California-Berkeley
School of Education
2030 Addison Street, #500
Berkeley, CA, 94720
Work: (510) 643-5206
Fax: (510) 642-2124
domsife@uclink4.berkeley.edu

Rowena Douglas
Ohio Department of Education
State Science Consultant
65 S. Front Street, Room 1009
Columbus, OH 43215
Work: (614) 466-2187
Fax: (614) 728-3058
pd_douglas@ode.ohio.gov

Brian Drayton
TERC
2067 Massachusetts Avenue
Cambridge, MA, 02140
Work: (617) 547-0430
Fax: (617) 349-3535
Brian-Drayton@TERC.edu

Sally Dudley
Dallas Independent School District
3700 Ross Avenue, Box 50
Dallas, TX 75204
Work: (214) 302-2449
Fax: (214) 302-2440
sdudley@tenet.edu

Jeane Dughi
Norfolk Public Schools
Department of Instruction
P.O. Box 1357
Norfolk, VA 23455
Work: (757) 441-2508
Fax: (757) 441-1589
jdughi@pen.k12.va.us

Hubert Dyasi
City College, CUNY
Education
Workshop Center
Room 4/220NAC
New York, NY 10031
Work: (212) 650-8436
Fax: (212) 650-6292
hdyasi@aol.com

Janice Earle
NSF / ESIE
4201 Wilson Boulevard
Arlington, VA, 22230
Work: (703) 306-1620
Fax: (703) 304-0412
jearlke@nsf.gov

Edgar Edwards
7802A Lemonwood Court
Richmond, VA 23228
Work: (804) 672-8518

Elizabeth Eisner
U.S. Department of Education
Planning & Evaluation Services
400 Maryland Avenue, SW
#6W210
Washington, DC 20202
Work: (202) 401-1857
Fax: (202) 401-3036
elizabeth_eisner@ed.gov

Jeanne Elliot
WestEd
Evaluation Research
730 Harrison Street
San Francisco, CA, 94107
Work: (415) 241-2734
Fax: (415) 512-2024
jelliot@wested.org

Sue Ellis
NASA Aerospace Education Services
Curriculum & Staff Development
1024 Silver Lake Boulevard
Frankfort, KY 40601
Work: (502) 695-6903
Fax: (205) 695-8542
sue@acsp.nasa.okstate.edu

Joni Falk
TERC
2067 Massachusetts Avenue
Cambridge, MA, 02140
Work: (617) 547-0430
Fax: (617) 547-3535
Joni_Falk@TERC.edu

Russell Faux
T E R C
Project Meet
2067 Massachusetts Avenue
Cambridge, MA 02140
Work: (617) 873-9613
Fax: (617) 349-3535
russell_faust@terc.edu

Alan Feldman
T E R C
2067 Massachusetts Avenue
Cambridge, MA, 02140
Work: (617) 873-9633
Fax: (617) 349-3535
alan_feldman@terc.edu

Joan Ferrini-Mundy
National Research Council
Center for Science, Math,
Engineering, & Education
2101 Constitution Avenue, NW
HA 450
Washington, DC 20418
Work: (202) 334-1467
Fax: (202) 334-1453
jferrini@nas.edu

John Finely
Atlanta Public Schools
Atlanta Systemic Initiative
P.O. Box 16925
Atlanta, GA 30321
Work: (404) 669-2250
Fax: (404) 669-2748

Bert Flugman
CUNY Graduate School
Center for Advanced Study In
Education
25 W. 43rd Street
New York, NY 10036
Work: (212) 642-2930
Fax: (212) 719-2488
bflugman@email.gc.cuny.edu

Judith Fonzi
University of Rochester
Warner Graduate School
3764 East Avenue
Rochester, NY 14618
Work: (716) 586-6050
Fax: (716) 586-0402
fonzij@aol.com

Gary Forde
St Louis Public Schools
Research Assessment &
Evaluation
911 Locust Street
St. Louis, MO, 63101
Work: (314) 231-3720
Fax: (314) 345-2648

Stanley Frankel
MESA Associates
6320 Bunchfield Ave.
Pittsburgh, PA 15217
Work: (412) 421-3421
Fax: (412) 421-9591
stanpburgh@aol.com

Joseph Frattaroli
Teachers Academy for
Mathematics & Science
Administration
3424 S. State Street
Chicago, IL, 60616
Work: (312) 949-2422
Fax: (312) 808-0103
jfrattaroli@tams.iit.edu

Linda Frazer
Department of Defense
Dependents Schools
Research & Evaluation
4040 N. Fairfax Drive
Arlington, VA 22203
Work: (703) 696-4471
Fax: (703) 696-8924
linda_frazer@odedodea.edu

Patricia Freitag
George Washington University
Educational Leadership
2134 G Street, NW
Washington, DC, 20052
Work: (202) 994-3567
Fax: (202) 994-5870
pfreitag@gwu.edu

Tom Gadsden
Eisenhower National
Clearinghouse
1929 Kenny Road
Columbus, OH 43210
Work: (614) 292-3330
Fax: (614) 292-2066
tgadsden@enc.org

Deborah Galloway
NASA Headquarters
Code FE
300 E Street
Washington, DC 20546
Work: (202) 358-1516
Fax: (202) 358-3048
debbie.galloway@hq.nasa.gov

Beth Gamse
ABT Associates, Inc.
55 Wheeler Street
Cambridge, MA 02138
Work: (617) 349-2808
Fax: (617) 349-2665
beth_gamse@abtassoc.com

Yolanda George
AAAS
Directorate for Education &
Human Resources Program
1200 New York Ave. NW
Washington, DC, 20005
Work: (202) 326-6670
Fax: (202) 371-9849
nbell@aaas.org

Roscoe Giles
Boston University
Center for Computational Science
3 Cummington Street
Boston, MA, 02215
Work: (617) 353-6078
Fax: (617) 353-6062
rosco@bu.edu

Alice Gill
American Federation of
Teachers
Educational Issues
555 New Jersey Avenue, NW
Washington, DC 20001
Work: (202) 393-6384
Fax: (202) 393-6371
agill@aft.org

Manuel Gomez
University of Puerto Rico
Resource Center for Science &
Engineering
P.O. Box 23334
San Juan, PR 00931-3334
Work: (787) 764-8369
Fax: (787) 751-9082
m_gomez@upri.upr.ciu.edu

Gerald Goumeau
Turtle Mountain Community
College
Tribal College Rural Systemic
Initiative
P.O. Box 340
Belcourt, ND 583 16
Work: (701) 477-8895
Fax: (701) 477-8896
jgomo@aol.com

Elizabeth Gray
Louisiana LACEPT
Mathematics
SLU - 10575
Hammond, LA 70402
Work: (504) 549-5897
Fax: (504) 549-2099
beth@selu.edu

Linda Griffin
Appalachian Rural Systemic
Initiative
200 W. Vine Street, #420
Lexington, KY, 40588
Work: (606) 255-3511
Fax: (606) 259-0986
lgriffin@arsi.org

Mary Gromko
Colorado Department of
Education
201 E. Colfax Ave.
Denver, CO 80203
Work: (303) 866-6764
Fax: (303) 866-6892
msgromko@iex.net

Ted Guffy
West Texas A&M University
Texas Rural Systemic Initiative
WT Box 60217
Canyon, TX 79016
Work: (806) 651-2603
Fax: (806) 651-2733
tguffy@mail.wtamu.edu

Wanda Guzman
Austin Independent School
District
Austin Collaborative for
Mathematics Education
1111 W. 6th Street
Austin, TX 78749
Work: (512) 414-4724
Fax: (512) 414-8360
wguzman@tenet.edu

Alfred Hall
Eisenhower Regional
Consortium for Mathematics &
Science Education
1700 N. Moore Street
Suite 1275
Arlington, VA 22209
Work: (703) 558-2247
Fax: (703) 276-0266
halla@ael.org

Eric Hamilton
NSF/EHR
4201 Wilson Boulevard, #875S
Arlington, VA 22230
Work: (703) 306-1682
Fax: (703) 306-0456
ehamilton@nsf.gov

Vivian Hampton
North Carolina A&T State
University
North Carolina AMP Program
College of Engineering
1601 East Market Street
Greensboro, NC 27411
Work: (336) 334-7447
Fax: (336) 334-7540
vivian@ncat.edu

Joseph Harris
The McKenzie Group
1100 17th Street, NW
Room 1100
Washington, DC 20036
Work: (202) 466-1111
Fax: (202) 466-3363
harrisj @mckgrp.com

Daniel Heck
University of Illinois
Department of Educational
Psychology
210 Education Building
MC 708
1310 S. Sixth
Champaign, IL 61820
Work: (217) 333-1450
Fax: (217) 244-0538
dheck@uiuc.edu

Mary Henry
Milwaukee Public Schools
Milwaukee Urban Systemic
Initiative
6620 W. Capitol Drive
Milwaukee, WI 53216
Work: (414) 438-3630
Fax: (414) 438-3470

Edward Hessler
Minnesota Department of
Children, Family & Learning
Science & Mathematics
1500 W. Highway 36
Roseville, MN 55113
Work: (651) 582-8792
ed.hessler@state.mn.us

Craig Hilmer
San Antonio USI
Evaluation
110 Tuleta
San Antonio, TX 78212
Work: (210) 734-0016
Fax: (210) 734-7890
chilmer@texas.net

Christopher Holle
Los Angeles Unified School
District
Los Angeles Systemic Reform
450 N. Grand Avenue
Room A-319
Los Angeles, CA 90012
Work: (213) 625-6421
Fax: (213) 626-7785

Ernest House
University of Colorado
School of Education
Campus Box 249
Boulder, CO 80309
Work: (303) 492-8863
ernie.house@colorado.edu

Michael Howard
West Virginia Department of
Education
Project CATS Internal Evaluator
729 Garvin Avenue
Charleston, WV, 25302
Work: (304) 346-4808
Fax: (304) 348-4808
howardm@ael.org

W. Jay Hughes
Georgia Southern University
Department of Sociology Center
for Rural Health & Research
PO Box 8148
Statesboro, GA 30460-8148
Work: (912) 618-0260
Fax: (912) 681-0816
jhughes@gasou.edu

Michael Hughes
Georgia Southern University
Curriculum, Foundations, &
Research
PO Box 8144
Statesboro, GA 30460-8144
Work: (912) 871-1554
Fax: (912) 681-5382
mahughes@gasou.edu

John Hunt
NSF / EHR
4201 Wilson Boulevard
Room 805N
Arlington, VA 22230
Work: (703) 306-1600
Fax: (703) 306-0399
jhunt@nsf.gov

Bill Hurt
Tennessee State University
Center of Excellence for
Research & Policy
330 10th Avenue, North
Box 141
Nashville, TN 37203
Work: (615) 963-7215
Fax: (615) 963-7214

David Imig
American Association of
Colleges for Teachers Education
1307 New York Ave, NW, # 300
Washington, DC 20005
Work: (202) 293-2450
Fax: (202) 457-8095
dimig@aacte.org

Kamil Jbeily
University of Texas Science
Education Center
Texas Regional Collaborative
1912 Speedway **SZB356**
Austin, TX, 78712
Work: (512) 471-9460
Fax: (512) 471-9244
kjbeily@mail.utexas.edu

Ehnima Johnson
NSF
4201 Wilson Boulevard, Rm 805
Arlington, VA 22230
Work: 7033061600
ejohnson@nsf.gov

Karen Johnston
NSF
Division of Undergraduate
Education
4201 Wilson Boulevard
Arlington, VA 22203
Work: (703) 306-1665 x5870
Fax: (703) 306-0862
kjohnsto@nsf.gov

Linda Jordan
Tennessee Department of
Education
5th Floor Andrew Johnson
Tower
7 10 James Robertson
Nashville, TN 37423-0937
Work: (615) 399-9209
Fax: (615) 532-8536
ljordan@mail.state.tn.us

Caroline Kaczala
Cleveland Heights-University
Heights City School
Assessment & Evaluation
2155 Miramar Blvd.
University Hts, OH 44118
Work: 2 1637 17434
Fax: 2163717177
c_kaczala@staff.chuh.org

Jane Butler Kahle
Miami University
Department of Teacher
Education
420 McGuffey Hall
Oxford, OH 45056
Work: (513) 529-1686
Fax: (513) 529-2110
kahlejb@muohio.edu

Robert Kansky
National Alliance of State
Science & Mathematics
Coalitions
11 Dupont Circle, NW
Suite 250
Washington, DC, 20036
Work: (202) 387-3600
Fax: (202) 387-4025
rkansky@nassmc.org

Joyce Kaser
Kaser & Associates
3301 Don Quixote, NW
Albuquerque, NM 87104
Work: (505) 7688833
Fax: (505) 764-8866
jskaser@aol.com

Conrad Katzenmeyer
NSF / REC
4201 Wilson Boulevard
Room 855
Arlington, VA 22230
Work: (703) 306- 1650
Fax: (703) 306-0434
ckatzenm@nsf.gov

Mark Kaufman
TERC
The Regional Alliance for Math
& Science
2067 Massachusetts Avenue
Cambridge, MA 02140
Work: (617) 873-9649
Fax: (617) 349-3535
mark_barnes@terc.edu

Judy Kelley
West Texas A&M University
Texas Rural Systemic Initiative
WT Box 60217
Canyon, TX 79016
Work: (806) 651-2271
Fax: (806) 651-2733
jkelley@mail.wtamu.edu

Anthony Kelly
NSF
REC/EHR
Stafford Building
4201 Wilson Boulevard, #855
Arlington, VA 22230
Work: (703) 306-1655
Fax: (703) 306-0434
aekelly@nsf.gov

Claudia Khourey-Bowers
Canton City Schools
Curriculum & Instructional
Grants
617 McKinley Avenue, SW
Canton, OH 44707
Work: (330) 438-2585
Fax: (330) 455-0682
claudia@raex.com

Jason Kim
Systemic Research
105 Eastern Avenue, #2 16
Dedham, MA 02026
Work: (781) 461-9021
Fax: (781) 461-9023
jkim@systemic.com

C. Eric Kirkland
COSMOS Corporation
3 Bethesda Metro Center, #950
Bethesda, MD 20814
Work: (301) 215-9100
Fax: (301) 215-6969
ekirkland@cosmoscorp.com

Mary Kopczynski
Urban Institute
State Policy Center
2100 M Street, NW
Washington, DC 20037
Work: (202) 261-5590
Fax: (202) 659-8985
mkopczyn@ui.urban.org

Kathryn Kozaitis
Georgia State University
Anthropology
University Plaza
Atlanta, GA 30303-3083
Work: (404) 651-1760
Fax: (404) 651-3235
antkxk@panther.gsu.edu

Carole Lacampagne
U.S. Department of Education
Postsecondary Institute
555 New Jersey Avenue, NW
Washington, DC 20208
Work: (202) 219-2064
Fax: (202) 501-3005
carole_lacampagne@ed.gov

Donna Landin
West Virginia Department of
Education
WVDE/IBM Reinventing
Education
1900 Kanawha Boulevard, East
Building 6, Room 346
Charleston, WV 25305
Work: (304) 558-0304
Fax: (304) 558-2584
dlandin@access.k12.wv.us

Terry Lashley
Appalachian Rural Systemic
Initiative
University of Tennessee
Knoxville, TN 37996
Work: (423) 974-4001
Fax: (423) 974-6436
tlashley@utk.edu

LeRoy Lee
Wisconsin Academy of
Sciences, Arts, & Letters
1922 University Avenue
Madison, WI 53705
Work: (608) 263-1692
Fax: (608) 265-3039

Okhee Lee
University of Miami
School of Education
University of Miami
Box 248065
Coral Gables, FL 33124
Work: (305) 284-5604
Fax: (305) 284-3003
olee@aol.com

Shelley Lee
State of Wisconsin
Department of Public Instruction
PO Box 7841
Madison, WI 53707-7841
Work: (608) 266-3319
Fax: (608) 264-9553
fishesa@mail.state.wi

Catherine Lewis
Developmental Studies Center
Research
240 Livoma Heights Road
Alamo, CA 94507
Work: (510) 533-0217 x271
Fax: (510) 464-3679
c_lewis@decstu.org

Marybeth Lima
Louisiana State University
Biological and Agricultural
Engineering
Room 149 E.B. Doran Building
Baton Rouge, LA 70803-4505
Work: (225) 388-1061
Fax: (225) 388-3492
mlima@gumbo.bae.lsu.edu

Jane Lindle
University of Kentucky
Center for the Study of
Education Leadership
111 Dickey Hall
Lexington, KY 40506-0017
Work: (606) 257-7845
Fax: (606) 323-9799
jclindle@pop.uky.edu

Gloria Lindner
Northern Arizona University
Science and Mathematics
Learning Center
P.O. Box 5697
Flagstaff, AZ 86011
Work: (520) 523-7160
Fax: (520) 523-7953
gloria.lindner@nau.edu

Madeleine Long
The Implementation Group
1420 New York Avenue, NW
Washington, DC 20005
Work: (202) 639-0671
Fax: (202) 639-0713
mlong@tig.vsadc.com

Julio Lopez-Ferrao
NSF / EHR
4201 Wilson Boulevard, #875S
Arlington, VA 22230
Work: (703) 306-1682
Fax: (703) 306-0456
jlopezfe@nsf.gov

Susan Loucks-Horsley
WestEd
4732 N. Oracle Road, # 217
Tucson, AZ 85705-1674
Work: (520) 888-2838
Fax: (520) 888-2621
sloucks@nas.edu

Sharon Lynch
George Washington University
Department of Teacher
Preparation & Special
Education
2134 G Street
Washington, DC 20052
Work: (202) 994-6174
Fax: (202) 994-3365
slynch@gwu.edu

James Madden
Louisiana State University
Mathematics
Lockett Hall
Baton Rouge, LA 70803
Work: (225) 388-1580
Fax: (225) 388-4276
madden@math.lsu.edu

Nancy Maihoff
 Delaware Department of
 Education
 Assessment & Accountability
 Townsend Building
 P.O. Box 1402
 Dover, DE 19903
 Work: (302) 739-2771
 Fax: (302) 739-3092
 nmahoff@state.de.us

Stephen Marble
 Southwest Education
 Development Laboratory
 211 East Seventh Street
 Austin, TX 78701
 Work: (512) 476-6861
 Fax: (512) 476-2286
 smarble@sedl.org

Cora Marrett
 University of Massachusetts-
 Amherst
 Whitmore Administration
 Building
 Amherst, MA 01002
 Work: (413) 545-2554
 Fax: (413) 545-2328
 cmarrett@provost.umass.edu

Wayne Martin
 ccsso
 State Education Assessment
 Center
 One Massachusetts Avenue, NW
 #700
 Washington, DC 20001-1431
 Work: (202) 336-7010
 Fax: (202) 789-1792
 waynem@ccsso.org

Robert Mathieu
 UW-Madison
 Astronomy
 475 N. Charter Street
 Madison, WI 53706
 Work: (608) 262-5679
 Fax: (608) 263-0361
 mathieu@astro.wisc.edu

Laurie Mathis
 Austin ISD
 Mathematics
 1111 W. 6th Street A-420
 Austin, TX 78703
 Work: (512) 414-4680
 Fax: (512) 414-8360
 lmathis@tenet.edu

Beverly Mattson
 RMC Research Corporation
 Regional III Comprehensive
 Center
 1815 N. Fort Meyer Drive, #800
 Arlington, VA 22202
 Work: (703) 558-4800
 Fax: (703) 558-4823

Kaye McCann
 Council For Basic Education
 Academic Standards Program
 1319 F Street, NW #900
 Washington, DC 20004-1152
 Work: (202) 347-4171
 Fax: (202) 347-5047
 kmccann@c-b-e.org

Angela McCormick
 Tennessee State University
 Center of Excellence for
 Research & Policy
 330 10th Avenue, N., Box 141
 Nashville, TN 37203
 Work: (615) 963-7232
 Fax: (615) 963-7214

Flora McMartin
 University of California-
 Berkeley
 NEEDS
 5935 Orchard Avenue
 Richmond, VA 94804
 Work: (510) 643-2928
 Fax: (510) 215-8281
 mcmartin@synthesis.org

Gene Meier
 Turtle Mountain Community
 College
 Tribal College Rural Systemic
 Initiative
 P.O. Box 340
 Belcourt, ND 58316
 Work: (307) 283-3110
 Fax: (307) 283-3110
 gmeier587@aol.com

F. Joseph Merlino
 LaSalle University
 Greater Philadelphia Secondary
 Math Project
 1900 West Olney Avenue
 Philadelphia, PA 19141
 Work: (215) 951-1203
 Fax: (215) 951-1843
 merlino@lasalle.edu

Robert Meyer
 WCER
 1025 W. Johnson Street
 767 Educational Sciences
 Madison, WI 53706
 Work: (608) 265-5663
 rhmeyer@aol.com

Susan Millar
 UW-Madison
 LEAD Center/NISE
 1402 University Avenue
 Madison, WI 53705
 Work: (608) 265-5943
 Fax: (608) 265-5923
 smillar@engr.wisc.edu

Vance Mills
 San Diego City Schools
 Mathematics & Science
 Department
 2441 Cardinal Lane
 San Diego, CA 92123
 Work: (619) 496-8127
 Fax: (619) 627-7373
 vmills@mail.sandi.net

Mary Mirabito
 Consulting for Human
 Resources
 Research & Development
 145 Avenue of the Americas
 #200
 New York, NY 10013
 Work: (212) 627-3988
 Fax: (212) 627-8733

Antoinette Mitchell
 The Urban Institute
 Education Policy Center
 2100 M Street, NW, #500
 Washington, DC 20037
 Work: (202) 261-5644
 Fax: (202) 833-2477
 amitchel@ui.urban.org

Donita Mitchell
 NCISE
 1726 M Street, NW, #704
 Washington, DC 20036
 Work: (202) 467-0652
 Fax: (202) 467-0659
 mitchell@ncise.org

Suzzane Mitchell
AR Statewide Systemic
Initiative
114 E. Capitol Avenue
Little Rock, AR 72201
Work: (501) 371-2062
Fax: (501) 371-2008
susunnem@adhe.arknet.edu

Gregory Moses
UW-Madison
1415 Engineering Drive
Madison, WI 53706
Work: (608) 263-1600
Fax: (608) 262-6400
moses@engr.wisc.edu

Robert Mount
Dallas Public Schools
School Planning & Evaluation
3709 Ross Avenue, Box 59
Dallas, TX 75204
Work: (214) 989-8454
Fax: (214) 989-8520.

Susan Mundry
Learning Innovations/WestEd
91 Montvale Avenue
Stonham, MA 02180
Work: (781) 279-8215
Fax: (781) 279-8220
smundry@wested.org

Michael Neuschatz
American Institute of Physics
Education & Employment
Statistics
One Physic Eclipse
College Park, MD 20740
Work: (301) 209-3077
Fax: (301) 209-0843
mneuscha@aip.org

John Nunnery
Memphis City Schools
Research, Standards &
Accountability
2597 Avery Avenue
Memphis, TN 38112
Work: (901) 325-5533
Fax: (901) 325-7635
nunneryj@memphis-schools.k12.tn.us

Ahmad Nurridin
NASA
Education Division
NASA Headquarters
Washington, DC 20546
Work: (202) 358-1517
Fax: (202) 358-3048
anurridin@hq.nasa.gov

Barbara Nye
Tennessee State University
Center of Excellence for
Research & Policy
330 10th Avenue, N., Box 141
Nashville, TN 37203
Work: (615) 963-7231
Fax: (615) 963-7214

Brian O'Callaghan
Southeastern Louisiana
University
Mathematics
500 Western Avenue
Hammond, LA 70402
Work: (504) 549-5894
Fax: (505) 549-2099
bocallaghan@selu.edu

Sandra O'Neal
Accountability and Development
Associates, Inc.
7200 Montgomery NE, # 228
Albuquerque, NM 87109
Work: (505) 898-6958
Fax: (505) 898-6914
onealsw@aol.com

Christine O'Sullivan
ETS
NAEP
P.O. Box 6710
Princeton, NJ 08541
Work: (609) 734-1918
Fax: (609) 734-1878
cosullivan@ets.org

Michael Oliver
Monterey Bay Heights Research
4249 Glen Haven Road
Soquel, CA 95073
Work: (831) 462-1181
Fax: (831) 465-1182
moliver@mail.telis.org

Maria-Carol Oriyomi
Atlanta Public Schools
Research, Planning, &
Accountability
222 Pryor Street, SW
Atlanta, GA 30341
Work: (404) 827-8091
Fax: (404) 827-8352
mcoriyoma@atlanta.k12.ga.us

Eric Osthoff
NISE
1025 W. Johnson Street
753G Educational Sciences
Madison, WI 53706
Work: (608) 263-5228
Fax: (608) 262-7428
erico@mail.wcer.wisc.edu

Irene Outlaw
San Diego City Schools
Mathematics & Science
Department
2441 Cardinal Lane
San Diego, CA 92133
Work: (619) 496-1814
Fax: (619) 627-7373
ioutlaw@mail.sandi.net

Lynette Padmore
Florida Collaborative for
Excellence in Teacher
Preparation
1540-G South Adams Street
Tallahassee, FL 32307
Work: (850) 561-2467
Fax: (850) 561-2684
lpadmore@fam.u.edu

Michael Palladino
University of Massachusetts
Donahue Institute
10 Tremont Street, 4th Floor
Boston, MA 02108
Work: (617) 367-8901
Fax: (617) 367-1434
palladino@umb.sky.cc.umd.edu

Michael Palmisano
Illinois Math & Science
Academy
Public Policy & Service
1500 West Sullivan Road
Aurora, IL 60506-1000
Work: (630) 907-5070
Fax: (630) 907-5940
mjp.imsa.edu

Hae_Seong Park
University of New Orleans
Educational Leadership
Counseling & Foundations
Lakefront
New Orleans, LA 70148
Work: (504) 280-6165
Fax: (504) 280-6453
hpark1@uno.edu

Edward Pauly
DeWitt Wallace Readers Digest
Fund
2 Park Avenue, 23rd Floor
New York, NY 10016
Work: (212) 251-9761
Fax: (212) 679-6990
epauly@wallacefunds.org

Terry Peard
Indiana University of PA
Biology
114 Weyandt Hall
Indiana, PA 15705
Work: (724) 357-2352
Fax: (724) 357-5700
tpeard@grove.iup.edu

Frances Pearlmuter
Horace Mann School
Science
231 W. 246th Street
Riverdale, NY 10471
Work: (718) 432-3962
Fax: (718) 548-2089
pearlmuter@horacemann.org

Kit Peixotto
Northwest Regional Education Lab
Mathematics & Science
Education Center
101 SW Main Street, #500
Portland, OR 97204
Work: (503) 275-9594
Fax: (503) 275-9584
peixottk@nwrel.org

David Perda
Massachusetts Department of
Education
Office of Math & Science
350 Main Street
Malden, MA 02148
Work: (781) 388-3300 x242
Fax: (781) 388-3395
dperda@doe.mass.edu

Dan Pike
LA Unified School District
Program Evaluation & Research
88 10 Emerson Avenue
Los Angeles, CA 90045
Work: (310) 215-9392
Fax: (310) 649-0926

Linda Plattner
Council For Basic Education
Academic Standards Program
1319 F Street, NW #900
Washington, DC 20004-1152
Work: (202) 347-4171
Fax: (202) 347-5047
lplattner@c-b-e.org

Andy Porter
NISE
1025 West Johnson Street
Madison, WI 53706
Work: (608) 263-4200
Fax: (608) 263-6448
acporter@mac.wisc.edu

Tracy Posnanski
University of Wisconsin-
Milwaukee
Center for Mathematics/Science
Education Research
2400 E. Hartford Avenue
Enderis Hall 265
Milwaukee, WI 53210-0413
Work: (414) 229-6646
Fax: (414) 229-4855
tjp@uwm.edu

Jennifer Presley
WestEd
1726 M Street, NW, #704
Washington, DC 20036
Work: (202) 467-0652
Fax: (202) 467-0659
presley@ncise.org

Jeffrey Priest
University of South Carolina
Ruth Patrick Science Center
Aiken Box 3
471 University Parkway
Aiken, SC 29801
Work: (803) 641-3269
Fax: (803) 641-3615
jeffp@aiken.sc.edu

Mike Puma
The Urban Institute
Education Policy Center
2100 M Street, NW
Washington, DC 20037
Work: (202) 261-5810
Fax: (202) 833-2477
mpuma@ui.urban.org

Kalyani Raghavan
University of Pittsburgh
ASSET Inc/LRDC
3939 O'Hara Street
Pittsburgh, PA 15260
Work: (412) 624-9580
Fax: (412) 624-9149
kalayani+@pitt.edu

Senta Raizen
NCISE
1726 M Street, NW, #704
Washington, DC 20036
Work: (202) 467-0652
Fax: (202) 467-0659
raizen@ncise.org

Linda Ramsey
Louisiana Tech University
School of Biological Science
P.O. Box 3179
Ruston, LA 71272
Work: (318) 257-4772
Fax: (318) 257-3852
lramsey@latech.edu

Jacqueline Raphael
The Urban Institute
Education Policy Center
2100 M Street, NW, #500
Washington, DC 20037
Work: (202) 261-5809
Fax: (202) 833-2477
jraphael@ui.urban.org

Caran Resciniti
Fresno Unified School District
3132 E. Fairmont
Fresno, CA 93726
Work: (559) 441-3642
Fax: (559) 265-2747
caresci@fresno.k12.ca.us

Peggy Richmond
Research & Evaluation
Associates
6320 Quadrangle Drive
Chapel Hill, NC 27514
Work: (919) 493-1661
Fax: (919) 489-0246

Gretchen Ridgeway
Department of Defense
Dependents School Act
Research & Evaluation
4040 N. Fairfax Drive
Arlington, VA 22203
Work: (703) 696-4471
Fax: (703) 696-8924
gretchen_ridgeway@ododedea.edu

James Ridgway
University of Durham
School of Education
Leazes Road
Durham, England DH1 1TA
Work: (011) 441-9137 x4353 7
Fax: (011) 441-9137 4350 6
jim.ridgway@durham.ac.uk

Liesel Ritchie
Mississippi State University
Social Science Research Center
P.O. Box 5287
Mississippi State, MS 39762
Work: (601) 325-0853
Fax: (601) 325-9062
liesel@lan.lssrc.msstate.edu

William Ritz
California State University
Science Education Department
1250 Bellflower Boulevard
Long Beach, CA 90840-4501
Work: (562) 987-1439
Fax: (562) 985-7164
writz@csulb.edu

Carol Robinson-Singer
Acct. & Development Assoc. Inc.
7200 Montgomery NE, #228
Albuquerque, NM 87109
Work: (505) 869-6105
Fax: (505) 869-7110
singercaro@aol.com

Steven Rogg
Illinois Mathematics & Science
Academy
Center @IMSA
1500 West Sullivan Road
Aurora, IL 60506
Work: (630) 907-5956
Fax: (630) 907-5946
rogg@IMSA.edu

Ingrid Rosemeyer
NISE
1025 West Johnson Street
Madison, WI 53706
Work: (608) 263-9250
Fax: (608) 262-7428
ibrosemeyer@facstaff.wisc.edu

James Rudolph
Boyer County Research
Services
9798 Townline Road
Petoskey, MI 49770
Work: (616) 347-0272
jrudolph@edcan.ehhs.cmich.edu

Wendy Russell
RMC Research Corporation
Regional III Comprehensive
Center
1815 N. Fort Myer Drive, #800
Arlington, VA 22202
Work: (703) 558-4800
Fax: (703) 558-4823

Daiyo Sawada
ACCEPT
Physics & Astronomy
Box 871504
Tempe, AZ 85287
Work: (602) 727-6799
Fax: (602) 727-6019
sawada@asu.edu

Kate Scantlebury
University of Delaware
Department of Chemistry and
Biochemistry
Newark, DE 19716
Work: (302) 831-4546
Fax: (302) 831-6315
kscantle@udel.edu

John Schoener
John Schoener & Associates, Inc.
P.O. Box 250378
Columbia University Station
New York, NY 10025
Work: (212) 666-3320
Fax: (212) 666-3325
jschoener@yahoo.com

Charles Shannon
St. Louis Public Schools
Evaluation & Research
911 Locust Street
St. Louis, MO 63101
Work: (314) 231-3720 x2347
Fax: (314) 241-0586
cshannon@dtl.slps.k12.mo.us

Patrick Shields
SRI International
Education & Health Division
333 Ravenswood
Menlo Park, CA 94025
Work: (415) 859-3503
Fax: (415) 859-2861
shields@sri.com

Jennifer Sidler
The Pew Charitable Trusts
Planning & Evaluation
2005 Market Street, #1700
Philadelphia, PA 19103
Work: (215) 575-4756
Fax: (215) 575-4888
jsidler@pewtrusts.com

Beth Skipper
Louisiana Systemic Initiatives
Program
Mathematics
1885 Wooddale Boulevard, 11th
Floor
Baton Rouge, LA 70806
Work: (225) 922-0690
Fax: (225) 922-0688
bskipper@regents.state.la.us

Ivan Small
Turtle Mountain Comm College
Tribal College Rural Systemic
Initiative
P.O. Box 340
Belcourt, ND 58316
Work: (406) 338-7379
Fax: (406) 338-7710
ivansmall@aol.com

Marshall Smith
U.S. Department of Education
400 Maryland Avenue, 7W300
Washington, DC 20202
Work: (202) 401-3389
Fax: (202) 401-0596
Mike_Smith@ed.gov

Barbara Spector
University of South Florida
15836 Sanctuary Drive
Tampa Palms, FL 33647
Work: (813) 971-1856
Fax: (813) 975-1015
spector@typhoon.coedu.usf.edu

Gerald Sroufe
AERA
1230 17th Street, NW
Washington, DC 20036
Work: (202) 223-9485
Fax: (202) 775-1824
jsroufe@gmu.edu

Mark St. John
Inverness Associates
15 Inverness Way
Inverness, CA 94937
Work: (415) 669-7156
Fax: (415) 669-7186
mstjohn@inverness-research.org

Bernice Stafford
The Lightspan Partnership, Inc.
School Marketing & Evaluation
10140 Campus Point Drive
San Diego, CA 92121-1520
Work: (619) 824-8309
Fax: (619) 824-8001
bstafford@lightspan.com

Bruce Stewart
University of California,
Berkeley
Lawrence Hall of Science
1 Centennial Drive
Berkeley, CA 94720
Work: (510) 643-7228
Fax: (510) 642-1055
bstew@uclink4.berkeley.edu

Craig Strang
University of California,
Berkeley
Lawrence Hall of Science
1 Centennial Drive
Berkeley, CA 94720
Work: (510) 642-9809
Fax: (510) 642-1055
cstrang@uclink.berkeley.edu

Jonathan Supovitz
University of Pennsylvania
Consortium for Policy Res in Ed
3440 Market Street #560
Philadelphia, PA 19104
Work: (215) 573-0700 x230
Fax: (215) 573-7914
jons@gse.upenn.edu

Larry Suter
Stanford University
School of Education
485 Lasuen Mall
Stanford, CA 94305
Work: (650) 725-7306
Fax: (703) 306-0434
lsuter@nsf.gov

Irene Swanson
Los Angeles Systemic Initiative
LA-SI Van Nuys MST
6625 Balboa Boulevard
Van Nuys, CA 91406
Work: (818) 997-2574
Fax: (818) 344-8379
iswanson@lausd.k12.ca.us

Bryan Szumlas
Calgary Catholic School District
Ascension of Our Lord School
509 Harvest Hill Drive, NE
Calgary, Alberta
Canada **T3K 4G9**
Work: (403) 226-5789
Fax: (403) 226-5798
bszumlas@telusplanet.net

Caroline Takemoto
Los Angeles Unified School
District
LA Systemic Initiative
450 N. Grand Avenue
Room A-3 19
Los Angeles, CA 90012
Work: (213) 625-6471
Fax: (213) 626-7785

David Taylor
Fayette County Public Schools
701 E. Main Street
Lexington, KY 40502
Work: (606) 281-0240
Fax: (606) 255-6901
dtaylor@fayette.k12.ky.us

Roberta Taylor
IBM Global Education
Consulting
818 Flagstone Lane, SE
Marietta, GA 30067
Work: (770) 321-0005
Fax: (770) 321-2651
bunnyt@us.ibm.com

Anthony Tezik
Department of Education
Division of Federal Programs
333 Market Street
Harrisburg, PA 17126-0333
Work: (717) 783-6903
Fax: (717) 783-6900
TTezik@education.state.PA.us

Tom Thompson
University of South Carolina
Educational Leadership &
Policies
College of Education
Columbia, SC 29208
Work: (803) 777-3091
Fax: (803) 777-3090
tthompson@ed.sc.edu

John Thorpe
National Council of Teachers of
Mathematics
1906 Association Drive
Reston, VA 20191
Work: (703) 620-9840
Fax: (703) 476-9027
jthorpe@nctm.org

Clara Tolbert
School District of Philadelphia
Office of Leadership & Learning
Administration Building
Room 705
Philadelphia, PA 19103
Work: (215) 299-7840
Fax: (215) 299-3472
ctolbert@phila.k12.pa.us

Jo Topps
California K-12 Alliance
431 **Havans** Avenue
Long Beach, CA 90814
Work: (562) 597-8523
Fax: (562) 597-6087
jo_topps@cams.edu

Jeff Turley
Arizona State University
Physics & Astronomy
ACEPT Project
P.O. Box 871504
Tempe, AZ 85281
Work: (602) 965-7903
Fax: (602) 727-6019
jtaccept@asu.edu

Naida Tushnet
WestEd
Evaluation Research
4665 **Lampson** Avenue
Los Alamitos, CA 90720
Work: (562) 985-9019
Fax: (562) 985-9635
ntushnet@wested.org

Mark Twiest
Indiana University of PA
Professional Studies in
Education
313 Davis Hall
Indiana, PA 15705
Work: 7243572400
Fax: 7243577515
mgtwiest@grove.iup.edu

Jerry Valdez
Fresno Unified School District
Instructional Services/ Science
Office
3132 E. Fairmont Avenue
Building 3
Fresno, CA 93726
Work: (209) 441-3684
Fax: (209) 265-2749
jdvsience@aol.com

Sally Valenzuela
San Antonio USI
110 Tuleta
San Antonio, TX 78212
Work: (210) 734-0016
Fax: (210) 734-7890
sara@tenet.edu

Roland VanOostveen
Toronto District School Board
Instruction Office
Level 2
Scarborough, Ontario
Canada M1P4N6
Work: (416) 396-7797
Fax: (416) 396-4292
rolan_vanoostveen@sbe.scarbor
ough.on.ca

Jean Vanski
NSF
Directorate for Education &
Human Resources
4201 Wilson Boulevard
Suite 805
Arlington, VA 22230
Work: (703) 306-1601
Fax: (703) 306-0399
jvanski@nsf.gov

Jeffrey Walczyk
Louisiana Tech University
Psychology
P.O. Box 10048
Ruston, LA 71272
Work: (318) 257-3004
Fax: (318) 257-2379
walczyk@latech.edu

Mary Walker
Austin Independent School
District
Science & Health Curriculum
111 W. 6th Street S.
Austin, TX 78703
Work: (512) 414-4662
Fax: (512) 414-1502
docwalk@tenet.edu

Jaclynn Walette
Turtle Mountain Community
College
Tribal College Rural Systemic
Initiative
P.O. Box 340
Belcourt, ND 58316
Work: (701) 277-1839
Fax: (701) 277-3600
nnylcaj@aol.com

Judy Walter
ASCD
Program Development
1703 N. Bearuegard Street
Alexandria, VA 22311
Work: (703) 575-5665
Fax: (703) 575-5985
jwalter@ascd.org

Patsy Wang-Iverson
Research for Better Schools
Mid-Atlantic Eisenhower
Consortium
444 N. Third Street
Philadelphia, PA 19123
Work: (215) 574-9300
Fax: (215) 574-0133
wang@rbs.org

Norman Webb
NISE
1025 West Johnson Street
Madison, WI 53706
Work: (608) 263-4287
Fax: (608) 263-6448
nlwebb@facstaff.wisc.edu

K. David Weidner
American Association of School
Administration
1801 North Moore Street
Arlington, VA 22209
Work: (703) 875-0760
Fax: (703) 807-1849
dweidner@aasa.org

Iris Weiss
Horizon Research, Inc.
111 Cloister Court, Suite 220
Chapel Hill, NC 27514
Work: (919) 489-1725
Fax: (919) 493-7589
iweiss@horizon-research.com

Jeffrey Weitz
Horace Mann School
231 W. 246th Street
Riverdale, NY 10471
Work: (718) 432-3992
Fax: (718) 548-2089
weitz@horacemann.org

Gerry Wheeler
National Science Teachers
Association
1840 Wilson Boulevard
Arlington, VA 22201
Work: (703) 312-9254
Fax: (703) 243-0407
gwheeler@nsta.org

Paula White
NISE
1025 W. Johnson Street
Madison, WI 53705
Work: (608) 263-4353
Fax: (608) 262-7428
pawhitel@facstaff.wisc.edu

Lois Williams
Baltimore City Public School
System
Office of Sci, Math & Health Ed
200 East North Avenue
Baltimore, MD 21202
Work: (410) 396-8585
Fax: (410) 396-8063
76522.1556@compuserve.com

Jonathan Wilson
Morgan State University
Baltimore Urban Systemic
Initiative
1700 East Coldspring Lane
Montebello, MD 21251
Work: (443) 885-3304
Fax: (410) 319-3324
jwilson@jewel.edu

Yuwadee Wongbundhit
Miami-Dade County Public
Schools
Division of USI Math & Science
1500 Biscayne Blvd., # 326B
Miami, FL 33132
Work: (305) 995-2923
Fax: (305) 995-2910
yuwadee@dcpa.dade.k12.fl.us

Robert Yin
COSMOS Corporation
3 Bethesda Metro Center, #950
Bethesda, MD 20814
Work: (301) 215-9100
Fax: (301) 215-6969

Ming Yu
LA Unified School District
Program Evaluation & Research
8810 Emerson Avenue
Los Angeles, CA 90045
Work: (310) 215-9392
Fax: (310) 649-0926
myu@lausd.k12.ca.us

Robert VanZant
San Diego City Schools
Mathematics & Science
2441 Cardinal Lane
San Diego, CA 92123
Work: (619) 496-1817
Fax: (619) 627-7373
bvanzant@mail.sandi.net

Andrew Zucker
SRI International
Social Sciences Department
1611 N Kent Street
Arlington, VA 22209-2173

Appendix B
Fourth Annual Forum Evaluation
Paula A. White, National Institute for Science Education

The Fourth Annual NISE Forum, *Evaluation of Systemic Reform in Mathematics and Science*, addressed important aspects of evaluating systemic reform: understanding evaluation of systemic reform, models and approaches to evaluation of systemic reform, and findings about systemic reform from evaluations and research. Panelists represented a wide range of expertise in systemic reform from across the nation. Participants were sent papers prepared by the panelists in advance of the Forum. Following each panel were small group discussions where participants analyzed and wrote responses to questions on the issues raised by the panel sessions. These discussion and networking opportunities aided the Forum in achieving its overall goal: to draw together leaders in the field and stimulate intellectually rich conversations to develop a better understanding of the evaluation of systemic reform.

Of the 270 participants, 34 percent represented professional organizations, 23 percent were education specialists from universities, 14 percent were district representatives, 10 percent were science, mathematics, and engineering specialists from universities, 8 percent were federal government officials, 5 percent were state representatives, 3 percent were Department of Education representatives, and 2 percent were corporation and foundation representatives (see Figure 1).

Organization Type	Percent
Professional organizations	34%
Education specialists from universities	23%
District representatives	14%
Federal government officials	11%
SMET specialists from universities	10%
State representatives	5%
Corporation and foundation representatives	2%
Total registrants: 270	

Figure 1. 1999 Forum Registrants, by Type of Organization

The Forum began with an overview of the NISE and charge for the conference from NISE Director Andrew Porter. NISE's Systemic Reform Co-Team Leader Norman Webb and John Hunt, Deputy Assistant Director of the Education and Human Resources Directorate of the NSF, then outlined their hopes for the conference. Questions on research and evaluation of systemic reform were raised during the three panel sessions.

In a wrap up session, Ernie House, Professor of Education at the University of Colorado, and Cora Marrett, Provost at the University of Massachusetts-Amherst, synthesized the panel presentations and the reports from the small group discussions, pulling together the work of a day and a half. House pointed out that no single evaluation design serves all purposes of systemic reform and, therefore, priorities need to be established on what evaluation should do, and model evaluations should be developed. Marrett stressed the importance of leadership on the evaluation front, if the promises of systemic reform are to be fulfilled and realized. Marshall Smith, Acting Deputy Secretary of the U.S. Department of Education and one of the original researchers of

systemic reform, followed these remarks with concluding statements regarding the nature and value of evaluating systemic reform.

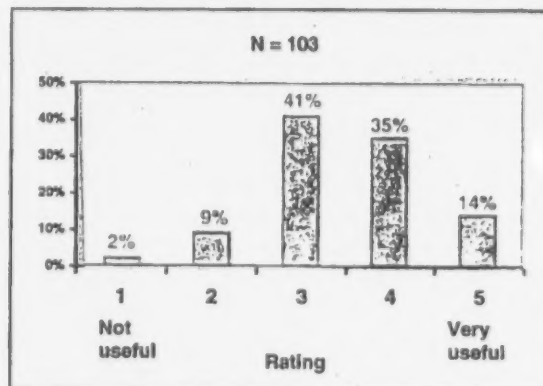
Summary of Responses

What follows is a summary of the evaluations completed by Forum participants following the third breakout session on the second day of the Forum. The response rate was 39 percent, with 104 of the 270 participants completing evaluations. Participants were asked to respond to the usefulness for them of the following aspects of the Forum:

- Panel I
- Panel II
- Panel III
- The Panelists' Papers
- The Small Group Discussions
- Other Opportunities for Networking
- The Forum Overall

Respondents were asked to rate each of these items on a scale from 1 to 5, with 1 signifying the lowest rating, "not useful," and 5 signifying the highest rating, "very useful."

How Useful Did You Find Panel I: Understanding Evaluation of Systemic Reform?



Note: Percentages are based on the total number responding to the question. Percentages may not total 100 due to rounding.

The average rating for Panel I was 3.50, with 90 percent giving Panel I a rating of 3 or higher. A few respondents wrote comments in addition to giving their rating. Positive responses made by participants about Panel I include:

Good overview and summary of systemic evaluation.

Dan Heck and Zoe Barley gave papers that offered important perspectives. Iris Weiss' talk inspired a sense of honesty and community for evaluators which I valued.

[The Panelists] raised points for good discussion.

All [the panelists] were extremely interesting from an intellectual aspect.

Being fairly new to the game of evaluation, I came away from this session feeling that evaluation of systemic reform is a daunting task, much more complex and complicated than my original perception.

Great to hear observations from experienced systemic reform evaluators.

Below are a few suggestions for improvement of Panel 1:

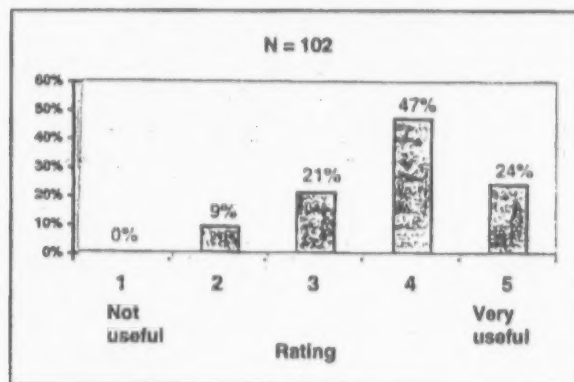
Would have liked to hear more detail on some of the methods mentioned.

More practical examples would have been appreciated.

Too much of a disconnect between the academic and practitioner perspectives.

It would have been more useful if the audience had participated.

How Useful Did You Find Panel II: Models and Approaches to Evaluation of Systemic Reform?



Note: All percentages are based on the total number responding to the question. Percentages may not total 100 due to rounding.

The overall average rating of Panel II was 3.85, with 92 percent giving Panel II a rating of 3 or higher. The following comments were made regarding the benefits of Panel II:

This panel kick-started a number of significant issues which led to significant discussion.

I liked the spark of contention and exchange among Norma Davila, Mark St. John, and Manuel Gomez.

It's always useful to see results. Mark St. John's comments brought up several conflicting opinions that were not resolved in small group discussion. Our group struggled intellectually; this was positive.

Excellent. Mark St. John's presentation was provocative.

Mark St. John provided enlightened clarity, transecting explorations of research to focus on what can be assessed, responsibly.

Norma Davila's data was useful to see how one SSI approached the issue of student achievement.

Excellent view of differing approaches to differing situations.

Jeanne Rose Century's idea of evaluation as advocate/supporter is very important.

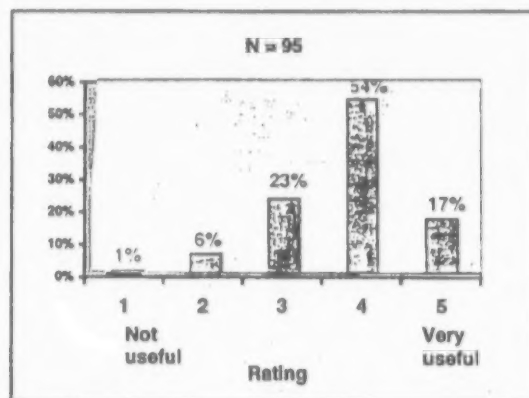
By the end of this session, I began to realize that the understanding of the evaluation of systemic reform is still developing and maybe the best way to approach the task is to just do it—but in a thoughtful, systemic way.

The following two suggestions were made about Panel II:

The presenters might have developed a common set of points/issues to focus around.

The models presented could have been more specific.

How Useful Did You Find Panel III: Findings about Systemic Reform from Evaluations and Research?



Note: All percentages are based on the total number responding to the question. Percentages may not total 100 due to rounding.

The average rating of Panel III was 3.82, with 94 percent giving Panel III a rating of 3 or higher. The following comments are a sampling by participants in response to Panel III:

*Informative, good examples and illustrations of concepts.
This was a most informative session!*

Many important issues were raised in this Panel, especially the need for new evaluation approaches appropriate to systemic reform.

Jane Butler Kahle, Daryl Chubin, and Robert Meyer had useful perspectives on evaluation and assessment. Chubin provided a strong responsible analysis of what research can provide now and in the future.

Most useful were ideas about data representation for powerful messages that need to be communicated succinctly to different audiences.

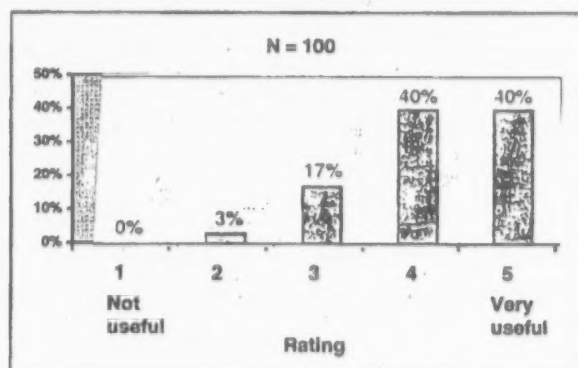
Given that the findings generally centered around student outcome oriented results, the session was very useful.

The questions and answers help bring the panel discussions to a higher level of effectiveness.

The following criticism was made about Panel III:

The thread helping the session hang together was not obvious to me. It seemed to be a collection of knowledgeable persons presenting their favorite methodologies and results, rather than focusing on the lessons that have been extracted.

How Useful Did You Find the Panelists' Papers?

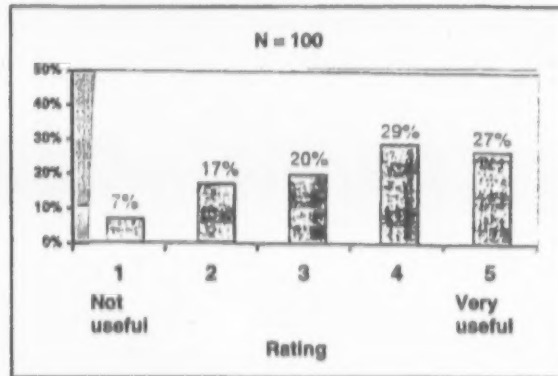


Note: All percentages are based on the total number responding to the question. Percentages may not total 100 due to rounding.

The average rating of the panelists' papers (sent to participants in advance of the Forum) was 4.17, with 97 percent giving the papers a rating of 3 or higher. The following is the only comment in response to the papers:

It was helpful to receive the papers in advance.

How Useful Did You Find the Small Group Discussions?



Note: All percentages are based on the total number responding to the question. Percentages may not total 100 due to rounding.

The average rating of the small group discussions was 3.52, with 76 percent giving the small group discussions a rating of 3 or higher. While the rating is still quite high, this item has the lowest overall rating. The comments on the small groups were quite diverse. The following are a sampling of comments by participants who responded very favorably to the small group discussions:

The best part of the Forum was the breakout group—mine left me with new ideas and information and plans.

The break out discussions were very productive and engaging.

Keeping us in the same groups [the three break out discussion groups] allowed us to quickly establish personal relations that encouraged deeper and more active participation.

The small groups were the best part. We had an excellent facilitator which lead to good discussions.

My colleagues and I felt the table discussions were very helpful.

The breakout sessions were great—very good conversation.

The following comments were made by participants who saw shortcomings in the small group discussions and had some suggestions to make:

Breakout session questions were not really conducive to discussion. I found them to be redundant and close-ended.

Small group sessions did not flow well. Questions were either too specific or redundant, causing discussions to drift off topic.

Perhaps assignment to groups should be a bit more purposive. Attend to balance of male/female, job role and participants' experiences/interests.

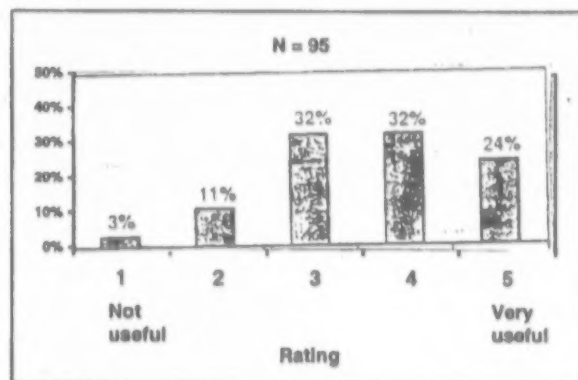
I would have preferred small groups identifying and discussing their own topics for discussion rather than adhering to a predetermined format and questions.

The small group discussions were too often the same people saying the same thing by the third time. Rotate groups.

The small groups would be improved by a closer and more dynamic link to the panels.

The broad diversity of the participants (prior knowledge and experience) had some positive contribution to the small groups, but overall limited the group's ability to come to consensus in the short time allowed.

How Useful Did You Find Other Opportunities for Networking?



Note: All percentages are based on the total number responding to the question. Percentages may not total 100 due to rounding.

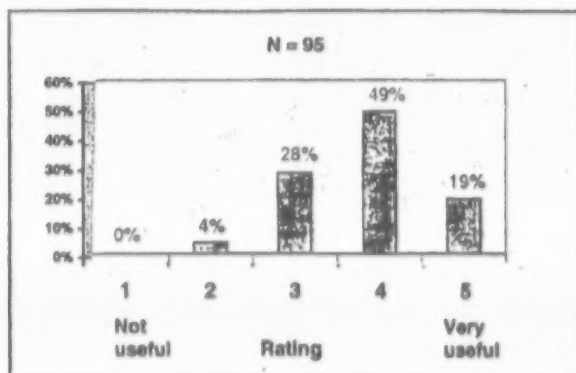
The average rating of other opportunities for networking was 3.66, with 88 percent giving a rating of 3 or higher. The following two suggestions were made by participants regarding opportunities for networking:

Provide more opportunities to network.

Use the reception as a time to allow for networking, not a presentation.

BEST COPY AVAILABLE

How Useful Did You Find the Forum Overall?



Note: All percentages are based on the total number responding to the question. Percentages may not total 100 due to rounding.

The average participant rating given to the Forum overall was 3.86, with 96 percent giving the Forum a rating of 3 or higher. Below are examples of participants' comments on the overall usefulness of the Forum:

I was fearful this would be a dull conference with presenters reading papers, but have been relieved to find presentations short, to the point, and very thought provoking. The format of presentation followed by small group discussion is very effective. Topics were also well chosen. Thanks for a good learning experience.

This Forum was a great opportunity to hear major contributors to the field and their latest thinking, as well as practitioners, take on what is happening and their thinking on the reform effort evaluation.

The Forum was good for an introduction to people and ideas and will guide my further study.

The Forum provided an opportunity to connect and network with colleagues.

Good mix of practitioners, evaluators, and policymakers.

Thank you for providing a detailed list of participants' contact information!

The Forum was a useful opportunity to gauge the state of the field and network.

Do You Have any Advice for Making the Forum More Useful?

Respondents provided valuable recommendations that will be considered in planning future NISE Forums. The recommendations are summarized below.

Panels

- The panels should be more interactive and challenge each other more rather than making sequential presentations.
- The presentations could be briefer (or fewer presenters) to allow more time for questions from the audience.
- Do not assume the moderator is best qualified to ask elucidating questions.
- Provide presenters with guidelines on overheads to ensure readability.
- Focus more on the practical than the theoretical, e.g., examples of evaluation plans could be presented and discussed to include lessons learned during implementation of the plan.
- Include a session on resources available such as technical assistance, on-line resources/conferences, and evaluating community and parent involvement.

Panelists ' Papers

- *Speakers should cover material other than what is in their papers.*
- *Speakers should avoid reading their papers.*

Small Group Discussions

- Rotate participants in small groups to allow for greater exposure to ideas.
- Use open-ended questions to promote quality discussion, e.g., In what direction do you think evaluation should be headed? Why?
- Use skilled facilitators.
- Develop a method to report back immediately from the breakout groups prior to the next panel session, e.g., if each group submitted one primary idea that emerged in its discussion, this could be translated onto overheads for a five minute presentation before beginning the next session.

Opportunities for Networking

- Provide more opportunities to network.
- Use the reception as a time to allow for networking, not a presentation.

Forum Overall

- Involve more practitioners.
- Involve more school district folks (evaluators). Include a school district speaker on the agenda.
- A glossary of terms would be useful.
- A poster session where attendees could display and discuss their work would be useful.
- More focus on systemic reform at the college level would be useful.
- Fewer large panel presentations and more opportunities for participants to find out from others what has worked and what hasn't.
- Forum facility should be more handicapped friendly.

Summary

The Fourth Annual NISE Forum provided a variety of opportunities for both formal and informal conversations about the evaluation of systemic reform. A summary of the evaluation responses indicates that participants valued the Forum and considered it useful. Overall, 96 percent of the respondents gave the Forum a rating of 3 or higher. The panelists' papers received the highest average rating of 4.17, with 97 percent rating the papers' usefulness at 3 or higher. Even the small group discussions, which received the lowest average rating of 3.52, were rated by 76 percent of the participants at 3 or higher.

The evaluation results indicate that, once again, the NISE hosted a successful Forum. Of the participants in the Third Annual Forum, 94 percent ranked it 3 or higher based on their overall gain; 89 percent ranked the Graduate Education Forum at 3 or higher on overall excellence, and 96 percent ranked the Fourth Annual Forum at 3 or higher on usefulness.

Single copy price is \$15.75. To order copies contact:

CENTER DOCUMENT SERVICE
Wisconsin Center for Education Research
1025 W. Johnson St., Room 242
Madison, WI 53706-1796
608/265-9698

NO PHONE ORDERS. PREPAYMENT REQUIRED FOR ORDERS UNDER \$20.00.

Price is subject to change without notice.

National Institute for Science Education
University of Wisconsin-Madison
1025 West Johnson Street
Madison, WI 53706

(608) 263-9250
(608) 262-7428 fax
niseinfo @ macc.wisc.edu
<http://www.nise.org>



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)



NOTICE

Reproduction Basis

X

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").